the theory of microeconomic policy

Nancy Smith Barrett

The Theory of Microeconomic Policy

The Theory of Microeconomic Policy

Nancy Smith Barrett
The American University

D. C. HEATH AND COMPANY
Lexington, Massachusetts Toronto London

Published simultaneously in Canada.

Printed in the United States of America.

International Standard Book Number: 0–669–83170–0

Library of Congress Catalog Card Number: 73–161

To My Mother and Father

PREFACE

Microeconomic theory in the United States has been heavily dominated by the laissez faire tradition of Adam Smith. The classical and neoclassical models of a capitalist economy that purport to demonstrate how the invisible hand of the market produces the most socially desirable allocation of resources have confused the debate over appropriate public policy actions. Although Keynesian economics has produced a generally accepted view that government has a role in regulating *aggregate* demand, there is still concern that government interference *within markets* has undesirable effects. Thus, despite the acknowledged breakdown of a market system in the face of external factors (air pollution, interdependent tastes), economies of scale, and the provision of public goods, microeconomic theory texts generally deal with how a market system operates rather than with what can be done to make it operate better.

In addition to avoiding policy issues the traditional approach is to begin with a discussion of the behavior of households and firms, which leaves most readers bewildered. Who cares what households and firms do? This book starts where most others end—with an examination of the underlying problem facing all economic systems: the need to allocate scarce resources as efficiently as possible. The discussion of resource allocation in chapters 1 and 2 is abstracted from any particular institutional context. In chapters 3 and 4 the market model and the Lange model of decentralized socialism are discussed and compared as institutional models that are supposed to produce efficient resource allocation. Once the allocative problem is put into perspective, the question of appropriate public policies follows naturally.

Later chapters provide the analytical framework for the study of resource allocation and the requisite tools for analyzing practical problems and policy issues. In addition to the standard neoclassical paradigm a number of contemporary topics are considered, including behavior under uncertainty, linear programming and duality, input-output analysis, behavioral theories of the firm, and stability analysis. An entire chapter is devoted to capital theory and resource allocation over time.

This book is designed for a course in microeconomic theory beyond the introductory level. No mathematics beyond elementary algebra is required.

Nancy Smith Barrett

CONTENTS

part **II**

TOOLS OF ANALYSIS

part **III**

part I

THE THEORY OF EFFICIENT RESOURCE ALLOCATION

1 Microeconomic Theory and Economic Policy

One sure way to start an argument among a group of economists is to ask for a definition of economics. Perhaps the only point of agreement would be that economics has to do with scarcity. If resources and hence the goods and services produced were unlimited, there would be nothing to economize. Therefore, it is the limitation of resources that gives rise to the "economic problem."

Some economists view economics as the theory of rational choice between mutually exclusive alternatives.[1] Others have defined it as the theory of optimization subject to certain limitations, or constraints.[2] Still others view economics as a study of the behavior of individuals seeking to maximize their material satisfaction given limited resources.[3]

1–1 BASIC PROPOSITIONS OF MICROECONOMICS

This book is about the allocation of economic, or scarce, resources. When resources are scarce, goods and services are scarce and society develops (consciously or unconsciously) a system to determine how to use the resources to produce goods and services, what goods and services to produce, and how to ration these goods and services among its members. Such a system is called an **economic system.** Although different societies may have different values, and although they may develop different sorts of institutions to achieve their objectives, there are some basic propositions that any society economizing scarce resources must accept.

1. *Scarce resources should not be left idle.*

Clearly, if resources are scarce, then involuntary unemployment of

[1] For instance, see Oskar Lange, *On the Economic Theory of Socialism* (Minneapolis: University of Minnesota Press, 1938), pp. 59ff., and Vivian Charles Walsh, *Introduction to Contemporary Microeconomics* (New York: McGraw-Hill, 1970).

[2] For instance, see Paul A. Samuelson, *Foundations of Economic Analysis* (Cambridge, Mass.: Harvard University Press, 1947).

[3] For instance, see Alfred Marshall, *Principles of Economics* (London: Macmillan and Co., Ltd., 1959). This book was first published in 1890.

these resources is wasteful. When scarce resources are not put to some use, then the economic system is not realizing its full productive potential. This does not imply that we all must work sixteen hours a day. Leisure as well as labor is scarce. Full utilization of labor resources implies that if a person *wants* to work (that is, if he prefers work to leisure at the prevailing wage), he should be able to find a job. Furthermore, useful capital equipment that is not technologically obsolete should not lie idle. However, this does not mean that all machines must be run continuously without shutdowns for maintenance. There is an optimal rate of utilization of capital equipment as well as labor. Machines that are run too hard may wear out much faster than those allowed shutdown periods. Consequently, full employment of resources means that the labor force and the capital stock should be used at some optimal rate, not necessarily that they be used to maximum capacity.

2. *The optimal technology for producing a good or service is that which uses the least resources.*

A technique that uses three man-days of labor is preferred to one using four man-days of labor when the outputs of the techniques are identical. Equivalently, from a given resource input, more is better than less. By the first technique, twelve man-days of labor can produce one and one-third times the output obtained by using the second technique.

Using resources in less productive techniques than are currently available is often called hidden unemployment. In our example, the use of a less productive technique gives the same output as if a more productive technique were used and one out of every four persons were involuntarily unemployed. If the more efficient technique were introduced into the industry, additional resources would be released to increase output in that industry or in other sectors of the economy.

3. *If someone can be made better off without making someone else worse off, the change would represent an improvement.*

Clearly, if some people can gain from a change without hurting others, the change would be considered an improvement by all concerned. Of course, those who have gained little or not at all may resent the greater gains by others as a result of the change. If such resentment actually makes these people worse off, then we cannot say that no one has been hurt by the change. In determining the effect of any change in resource allocation on the welfare of individuals, all interactions must be accounted for. Modern economic theory recognizes that individual welfare may depend as much upon relative levels of material consumption as upon absolute levels.

An individual society will have other economic goals. Some countries may be concerned with equitable distribution of income, others with rapid growth or the development of a strong military establishment. But the criteria just listed are consistent with any of these specific objectives. These criteria must be observed if scarce resources are to put to their best use, however that concept is defined.

1-2 PERSPECTIVES OF MICROECONOMIC THEORY

Microeconomics, the study of resource allocation, has three major perspectives:

1. *Microeconomics describes the behavior of economic units.*

By analyzing how individual households and firms make choices in the face of scarcity, microeconomics can predict how they will react to changes in conditions that affect their choice between economic alternatives. How will an excise tax on automobiles affect car sales? Will manufacturers pass on a sales tax to consumers? How will a rising trend in real wages affect labor force participation of married women? What fiscal measures would cause polluting industries to switch to cleaner processes?

2. *Microeconomics develops criteria for optimal resource allocation.*

Given certain objectives with known constraints (or scarcities), what are the rules that must be followed if the optimum is to be achieved? Such an analysis has implications for any decision maker concerned with resource allocation — in business, government, a large corporation, or the central planning board of a socialist state. A labor union trying to maximize incomes of its members, an airline that wants to minimize the cost of flying a particular schedule, a household that wants to get the most nutrition from its food budget can all utilize the principles of microeconomic theory.

3. *Microeconomics can be used to evaluate behavior and economic institutions.*

Although historically microeconomics has been developed largely in reference to a market system, its concepts may easily be applied in discussing the problems of resource allocation in any society. Microeconomics is useful for evaluating alternative institutions and economic policies that a society might devise for achieving optimal resource allocation. For instance, what is the best way to achieve efficient urban transportation? Should college education be financed by individuals or the state? How can we deal most effectively with the massive economies of scale that characterize modern industrial technology?

1-3 THE LAISSEZ FAIRE TRADITION

An approach that combines an exposition of microeconomic theory with matters of economic policy may, at first glance, appear to be a misguided effort, flying in the face of the laissez faire tradition that has pervaded economics for two centuries. The idea that a free market with a minimum of interference from government is the best mode of resource allocation rests largely on the doctrine of the invisible hand set forth by Adam Smith in 1776 in *The Wealth of Nations*. Smith stated:

> Every individual is continually exerting himself to find out the most advantageous employment for whatever capital he can command. It is his own

> advantage, indeed, and not that of the society, which he has in view. But the study of his own advantage naturally, or rather necessarily, leads him to prefer that employment which is most advantageous to the society.[4]

The notion that individuals acting in self-interest achieve an optimal allocation of economic (scarce) resources from the point of view of society as a whole led naturally to the conclusion that no outside arbiter such as the central government should interfere with individual choices. This resulted in the enunciation of the laissez faire principle:

> No regulation of commerce can increase the quantity of industry in any society beyond what its capital can maintain. It can only divert a part of it into a direction into which it might not otherwise have gone; and it is by no means certain that this artificial direction is likely to be more advantageous to the society than that into which it would have gone of its own accord.[5]

Economic theory since Smith has largely been concerned with showing the conditions under which a laissez faire policy (or more appropriately, nonpolicy) will result in a socially optimal allocation of resources. Economic policy has, for the most part, been aimed at establishing these preconditions. Antitrust legislation, advocacy of free trade, measures to improve labor and capital mobility are all examples of economic policies consistent with the laissez faire approach to resource allocation. That is, these policies are designed to structure economic institutions so that optimal resource allocation will occur in the absence of direct government intervention. Some classical liberals like Milton Friedman have gone so far as to advocate the abolition of minimum wage laws and provisions prohibiting job discrimination.[6]

However, economists have also recognized areas in which a free enterprise system fails to provide the most desirable allocation of resources. Air pollution, poverty, excessive concentrations of economic power (and hence political power) in huge corporations and labor unions are all examples of some undesirable results of laissez faire attitudes. Yet the laissez faire tradition dominates thinking in the area of resource allocation. The antitrust division of the Department of Justice still prosecutes monopolies with a vengeance, on the assumption that breaking a single corporate giant into two or three corporate giants will somehow benefit consumers. Much of the controversy over wage-price controls to alleviate inflation centers on whether such controls would interfere with "free market forces."

Such policies and controversies are anomalies indeed, pursued as they are in an environment that is quite obviously in violation of the preconditions set forth by traditional economic theory for optimal resource allocation. Not only is American industry not perfectly competitive but economic theory cites many instances

---

[4] Adam Smith, *The Wealth of Nations* (New York: Random House, 1937), p. 421.

[5] Ibid.

[6] Milton Friedman, *Capitalism and Freedom* (Chicago: University of Chicago Press, 1962), chapter 7.

in which even a perfectly competitive economy will not produce a socially optimal resource allocation. Thus, although economic theory leads us to support laissez faire policies in some instances, it can also provide a guide for economic policy when laissez faire is clearly not advantageous.

Although economic theory has been used to justify free market capitalism, it can be shown that there are alternative economic systems that can also produce a desirable allocation of resources. Central planning, for instance, can sometimes be designed to achieve what a market system fails to do. Various economic systems have their strengths and weaknesses, which can be evaluated, in part, using the principles of microeconomic theory.

1–4 MICROECONOMIC THEORY AND THE KEYNESIAN REVOLUTION

One of the most obvious historical failures of laissez faire to achieve optimal resource allocation was the Great Depression of the 1930s. The Keynesian revolution in economic theory, which followed on the heels of that depression, clearly established a role for a public economic policy other than laissez faire. After the publication of Keynes's *The General Theory of Employment, Interest, and Money* it became acceptable for government to intervene in the economy to promote full employment. This intervention, however, was to be strictly on the "aggregate" level, that is, it was not to interfere with the market mechanism. However, it is unlikely that the public sector can be expected to expropriate ever larger portions of the national product in the name of functional finance or economic stabilization without interfering with the allocation of resources by the private sector. Furthermore, in the United States, as public spending rose from 10 per cent of GNP in 1929 to 31 per cent in 1969, the allocation of an increasing portion of the nation's resources was taking place outside the market system. Although Keynes foresaw the inevitability of the public sector growing more rapidly than the economy as a whole, he failed to consider the consequences for the allocation of resources among alternative uses. During the Great Depression, when he wrote *General Theory,* a substantial portion of the nation's resources were unemployed and any enterprise, regardless of its social worth, was preferable to leaving those resources idle.

> . . . if the Treasury were to fill old bottles with banknotes, bury them at suitable depths in disused coal mines which are then filled up to the surface with town rubbish, and leave it to private enterprise on well-tried principles of *laissez-faire* to dig the notes up again . . . , there need be no more unemployment and, with the help of the repercussions, the real income of the community and its capital wealth also, would probably become a good deal greater than it actually is.[7]

[7] John Maynard Keynes, *The General Theory of Employment, Interest, and Money* (New York: Harcourt Brace Jovanovich, 1965), p. 129.

He also suggested that endeavors of little use value, such as pyramid building and the erection of cathedrals, would be sure to generate full employment, since the populace would never complain of having too many of them. "Two pyramids, two masses for the dead, are twice as good as one; but not so two railways from London to York."[8]

One thing that Keynes failed to recognize is the irreversibility of public spending programs once they are initiated. The idea that spending for the sake of spending is a good thing presents serious questions for resource allocation in a full employment economy. What happens when the Office of Pyramid Construction develops such economic and political power that it wins increasingly large budgetary appropriations from the Congress at the expense of education, medical care, and other endeavors with obvious social merit?[9]

Basically, however, in establishing an active role for government in the economy, Keynesian economics has resulted in nearly one third of United States GNP being purchased by the government. Decisions concerning what, how, and for whom goods and services are produced are made outside a market system.

An analogy can be made with the large corporation—the dominant type of producer in the manufacturing sector of the American economy in terms of value added in production. These corporations often produce their own input (or dominate the sources of supply), service their own products, and control large pools of specialized labor. The decisions concerning how to allocate resources among alternative product lines must often be made in the absence of market criteria such as prices.

Lange once noted that optimal decisions cannot be made without knowing the terms on which alternatives are offered.[10] This proposition is the key to understanding how microeconomic theory can guide optimal decision making in the absence of a market. Put in another way, the laissez faire principle states that under certain conditions a free market system (with no government interference) produces an optimal allocation of resources. However, microeconomics is a way to analyze economic options in the absence of market information. Furthermore, it can indicate when a free market system is likely to fail to provide the optimal institutional framework.

1–5 MICROECONOMIC POLICY IN A MARKET SYSTEM

Microeconomic policy is important for a market system in two types of circumstances. The first is the case of market failures, that is, when a free market economy fails to produce optimal resource allocation. The second is where large amounts of goods and services are allocated outside markets, that is, within and among large corporations and in the public sector.

[8] Ibid., p. 131.

[9] For further discussion see Nancy S. Barrett, *The Theory of Macroeconomic Policy* (Englewood Cliffs, N.J.: Prentice-Hall, 1972), pp. 338–339.

[10] Lange, *Economic Theory of Socialism*, p. 60.

From the time of Adam Smith traditional economic theory has recognized that there are certain areas in which private decisions will not produce socially desirable results. There are many instances where private interests differ from social interests. My decision to plant a garden will be based only on the satisfaction I expect to receive from the garden. But this private calculation will understate the benefit derived by the entire neighborhood from such an improvement. A petroleum refining company emitting sulfur oxides does not include such emissions in a calculation of its production costs, even though air pollution represents a *social* cost. Only to the extent that private and social costs and benefits coincide will laissez faire result in maximum social well-being.

Another area in which a laissez faire policy fails to solve society's economic problems is income distribution. Economists have traditionally treated the question of optimal income distribution as a normative one, that is, requiring individual value judgments. In other words, the question of whom goods should be produced for is regarded as having no scientific answer. However, it may be apparent that the distributive consequences of some particular action are not beneficial from the viewpoint of the bulk of society. Consider the case of a wealthy recluse living in the midst of a poor community. His estate lies in the path of the only access road to the market, and he exacts a toll from all those who use the road. Suppose the tolls are regulated by a government agency. Noting that the roads are congested, the recluse requests an increase in the tolls. His argument is that if the tolls rise, fewer people would use the road and consequently it would be less congested. However, the agency in making its decision would certainly take the distributive consequences of such an action into account. An alternative might be to build a new road from revenues obtained by taxing the property of the wealthy recluse. In both cases, the allocative effect would be to reduce the traffic congestion and hence the cost of transportation to and from the market. The distributive consequences, however, would be quite different.

Finally, microeconomic policy is needed in a market economy when allocative decisions are made outside the market system. The growth of the public sector because of the growing demand for public goods and the acceptance of Keynesian principles of functional finance, and the growth of large corporations because of economies of scale and other advantages associated with size, has meant that many decisions concerning what, how, and for whom goods and services should be produced are made by people other than autonomous market participants. Microeconomic theory can provide criteria for resource allocation under those circumstances.

1–6 MICROECONOMIC POLICY INSTEAD OF A MARKET SYSTEM

Microeconomic theory can provide criteria for optimal resource allocation. Although the economic objectives of a society are related to its institutions, the notion of optimal resource allocation can be considered apart from the institu-

tions a society develops to achieve its goals. Thus, given social objectives, microeconomics can be used to evaluate social institutions and policies designed to achieve those objectives. Furthermore, given certain objectives, theory would suggest the best policies to achieve those ends, with laissez faire being one possibility.

One example of how microeconomic theory can be used for economic decision making in a nonmarket setting is **benefit-cost analysis,** now widely used in government and industry. In many such studies values, or **shadow prices,** must be imputed to resources when market values are either not available or else assumed to be irrelevant because of the divergence of private and social returns. For instance, in determining an optimal rate of government investment, some economists advocate use of an interest rate (discount factor) that is lower than the market rate. This is because it is expected that society as a whole has a weaker preference for current consumption than individuals, because of the limited life-span of the latter. Thus society, including future generations, would generally prefer longer-lived investments and a higher level of investment (with less current consumption) than individuals of the present generation.

1–7 METHODOLOGICAL FRAMEWORK: ECONOMIC MODELS[11]

Economic theory involves formulating and describing hypotheses about economic behavior, recognizing interrelations among various aspects of economic behavior, and analyzing the implications of these interrelations. One way to examine the hypotheses and interrelations of economic theory is to construct a **model.**

A model distills the essence of the problem under investigation. It simplifies the analysis by stripping away the unnecessary assumptions and empirical details while retaining those needed to obtain useful insights and operational results. A test of the skill of a theoretician is his ability to simplify a complex problem.

Suppose we are interested in how price is determined in a particular market. A model of the market can be based upon certain hypotheses about market behavior as well as upon the interrelations between buyers and sellers in the market. Suppose we assume

1. The quantity of the good demanded is a linear function of its price, or algebraically,

$$q_D = a_1 + a_2 p$$

[11] The material in this section is largely taken from the introductory chapter of Barrett, *The Theory of Macroeconomic Policy.*

2. The quantity of the good supplied is also a linear function of its price, or

$$q_S = b_1 + b_2p$$

3. Buyers and sellers will adjust their offering prices until the market is cleared, that is, until quantity demanded equals quantity supplied, or

$$q_D = q_S$$

The first two equations are **behavioral relations;** the third is an **equilibrium condition.**

A model must not only be simple but must also yield operational results, that is, it must be capable of providing answers to relevant questions. Suppose we want to predict market price on the basis of our model.

The complete model

$$\begin{aligned} q_D &= a_1 + a_2p \\ q_S &= b_1 + b_2p \\ q_D &= q_S \end{aligned}$$

has three equations and three unknowns. Consequently, it is consistent, that is, the equilibrium market price can be determined (except in special cases).[12]

In equilibrium

$$q_D = q_S$$

and substituting the behavioral relations for q_D and q_S, the equilibrium condition can be expressed as

$$a_1 + a_2p = b_1 + b_2p$$

Solving for p, the price in equilibrium is found to be

$$p = \frac{b_1 - a_1}{a_2 - b_2}$$

A model may be used to answer other questions. How does the market price fluctuate over time? What happens when the system is not in equilibrium? How would market price respond to an increase in demand? To answer these questions more information may be needed. However, the answers can all be found in the context of the model just presented.

Mathematics in Economic Models

A model of economic theory need not be mathematical; however, mathematics has several attributes that contribute to its usefulness in economic theorizing. First,

[12] The results should also be economically meaningful. In this case, equilibrium price and quantities demanded and supplied should be non-negative.

it provides a shorthand notation in which to express the variables of the model. Then, mathematics is a powerful tool of analysis that can contribute to problem solving in economics. Differential calculus, for instance, can provide techniques for optimization analysis, so important to questions of resource allocation.

Although mathematics is not the only tool for model building, it is an important one. However, it can only be used once the underlying economic relations of the model have been specified. The specification of a model, the development of behavioral hypotheses, and recognition of underlying relations are clearly the most challenging aspects of economic theory.

QUESTIONS FOR STUDY AND REVIEW

1. How can you explain the traditional association of microeconomic theory with laissez faire capitalism? Describe some other areas in which microeconomic theory can provide fruitful insights for economic policy.
2. What is the relevance of Keynesian (macro)economics to microeconomics? Of microeconomics to Keynesian (macroeconomic) theory?
3. "During World War II, resource allocation in the United States was at an all time optimum, since both the labor force and capital stock were being utilized to their maximum capacity." Discuss.
4. Discuss the model-building approach as a methodology in economic theory. Explain the statement in the text that the specification of a model is clearly the most challenging aspect of economic theory.

ADDITIONAL READING

Boulding, Kenneth E. "Economics as a Moral Science," *American Economic Review, LIX* (March 1969), pp. 1–12.

Friedman, Milton. *Essays in Positive Economics*. Chicago: University of Chicago Press, 1953, part I.

Heilbroner, Robert. *The Worldly Philosophers*. New York: Simon and Schuster, 1967, chapter 3.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1970.

2 Resource Allocation and Economic Efficiency

Scarcity is the basic problem facing economic science. Since resources and thus goods and services are scarce, it is clear that full utilization of these resources is an important goal of an economic system. However, full utilization alone does not ensure that resources are being used to their best advantage. A person may work fifteen hours per day making square wheels, while another works eight hours producing round ones. From the viewpoint of society, the use of the labor force resource producing round wheels for eight hours may be more beneficial than the fifteen-hour effort that produces square ones. Thus, even if resources are fully utilized, the question of *what* is produced is certainly relevant to the problem of scarcity.

Suppose a man spends eight hours producing a useful item such as round wheels, but a machine is available with which he can produce the same number of wheels in an hour. The labor-intensive method may prove to be a waste of resources, depending, of course, on the cost of the machine. Clearly, the question of *how* resources are used to produce goods and services is relevant to the problem of scarcity.

Finally, suppose there are two wagon makers in the economy. One has an excessive inventory of wheels, while the other has halted production because of a lack of wheels. Now suppose the wheel maker delivers his wheels to the wagon maker with an excessive inventory of wheels. Although he has expended his labor (a scarce resource) on a useful product, using an efficient technique, he has not lessened scarcity, since he has delivered his goods to a person who already has an abundance of that product. Consequently, the question of *for whom* goods and services are produced is also relevant to the question of scarcity.

Optimal resource allocation clearly involves more than the assurance that all resources are fully employed. They must also be used efficiently to produce useful goods and services that are distributed to individuals who want them. In this chapter we will consider the criteria economists use to determine when, in fact, optimal allocation of resources has been achieved. In other words, how does one know that scarce resources have been used to best advantage?

2-1 A DIGRESSION ON SCARCITY

Although the basic economic resources, labor, capital, and natural resources are available in limited amounts, the rates at which they are utilized can be varied. The notion of full employment is consistent with many rates of resource utilization, and a resource may be viewed as scarce even though it is not being used at its maximum potential.

The labor of each individual, for instance, is potentially available for twenty-four hours a day. However, use at that rate would be shortsighted, since human beings need rest, food, and recreation to maintain labor effectiveness for any length of time. Furthermore, individuals may prefer to consume leisure in the place of goods and services. Even though leisure is not technically produced, it uses resources (since it holds labor resources from production) and consequently must be viewed as an economic, or scarce, good. Thus resources voluntarily allocated to leisure cannot be viewed as unemployed.[1] Maximum labor force participation does not necessarily represent optimal resource allocation, since there may be inadequate allocation of resources to leisure.

Labor resources can be increased in ways other than reducing allocation to leisure. Women, children, and the elderly can all be called into service. Although this practice has been followed in times of national emergency, it does not necessarily represent an improved allocation of resources if individuals prefer to use their resources in endeavors such as child rearing, education, and recreation. It is a mistake to consider such activities as out of the sphere of economics, because they all involve the use of scarce resources. The allocation of time to activities outside the labor force is an important consideration for economic theory. As goods and services become increasingly abundant, it is likely that individuals will desire to spend less time producing them and more time in other activities. Consequently, this area of economic analysis is likely to increase in importance.

We have seen that the notion of full employment of labor must be viewed in terms of all potential uses of that resource. In addition, all labor force activity does not involve direct production of consumer goods and services. Some labor resources are transformed into another resource, capital, which is then used to produce other goods and services. Capital is a produced means of production. The question of how much labor to devote to producing capital is an important problem of resource allocation. Furthermore, the notion of full employment of capital is not as straightforward as it might seem.

General Electric invested millions of dollars in capital equipment to manufacture black and white television. With the advent of color television, some of the equipment was converted but much of it had to be scrapped and replaced with completely new machines. Was the scrapping of machinery that had not

---

[1] If workers are involuntarily unemployed, their consumption of leisure is involuntary and leisure, like excessive wheel inventories, is no longer a scarce commodity for those individuals.

outlived its productive service an uneconomic decision? Not necessarily. The optimal rate of replacement of machinery must be viewed in terms of technological advance as well as in relation to changes in demand for products. If demand for black and white television decreases, or if machines that can produce both black and white and color components are available, then machinery for producing only black and white components becomes a redundant resource, even though the labor resources that produced it are scarce. This does not violate the principle that scarce resources must be fully employed if they are to be allocated to the best advantage.

Finally, natural resources can be utilized at different rates without compromising the full employment principle. Although the total stock of many of the natural resources available to us and to future generations is fixed, the flow rate of utilization affects the rate of depletion. The trade-off between immediate and future use of a fixed stock of resources is again a question of resource allocation. For instance, many economists believe that governments should control the rate at which the present generation depletes natural resources, since the limited horizons of individuals undervalue future resource requirements. If these economists are correct, this is a case where laissez faire breaks down as means of optimal resource allocation because of the divergence of private interests (of the present generation) and the social interest (including that of future generations).

Thus full employment of resources cannot be defined in any one way. In fact, the question of the optimal rate of resource utilization or supply is one of the important issues of microeconomic theory and policy. We will return to that question in this and later chapters.

2-2 CRITERIA FOR OPTIMAL RESOURCE ALLOCATION

Designing a policy to produce optimal resource allocation presupposes certain social goals or objectives. Criteria by which alternative allocations can be ranked or evaluated are called **welfare criteria.** Welfare criteria can be used to evaluate alternative ways of raising tax revenue, of transportating coal from Newcastle to London, of abating air pollution, and so on.

Since the concern of economic analysis from the time of Adam Smith has been with the question of optimal resource utilization, the earliest welfare criteria date from classical welfare economics. Then, at the beginning of this century, Vilfredo Pareto developed what has come to be known as the New Welfare Economics. Since then additional welfare criteria have been suggested.

The Classical View of Social Welfare

The classical economists held a somewhat organic view of society, based on the notion that social welfare is greatest when all the individuals in society, as

a group, are happiest. The satisfaction each individual gains from consuming goods and services is a psychic entity that the utilitarian philosophers called **utility.** Utility was viewed as a measurable amount of the psychic gratification received from consuming goods and services. Conceptually, a utility meter could be devised and strapped to the head of a consumer to register the intensity with which he or she enjoys certain goods and services. It was assumed that utility could be compared between individuals and between different types of goods. The utility I gain from watching a movie could be compared to that you enjoy when eating an orange.

Since social welfare is maximized when the sum total of the utilities for all individuals is the greatest, it would appear that the greatest social welfare would result from distributing goods in accordance with individuals' hedonistic capacities. Those with the greatest capacity for pleasure would receive the lion's share of the national product. However, the classical view was considerably more egalitarian. It assumed that all individuals have the same capacity for enjoying goods and services. But, for any individual, as more goods are consumed, the additional utility associated with consuming even more diminishes. Thus at high income levels gains in income will add less to utility than at low income levels.

The assumption that all individuals have the same capacity for enjoying goods and services, coupled with the proposition that gains in income produce less utility at high income levels than at low levels of income, led classical welfare economists to conclude that the distribution of income that would maximize social welfare is perfect equality. Suppose the individuals in an economy had the identical consumption utility functions shown in Table 2–1. **Marginal utility** (*MU*) is the increment to total utility associated with an additional $1,000 consumption of goods and services. As consumption increases, marginal utility declines. Suppose the economy consists of three families who consume all their income. GNP is $6,000. Table 2–2 shows three alternative distributions of income. In case I one family receives all the income and total utility is 350. As

TABLE 2–1
Utility as a Function of Total Consumption

Consumption (Dollars)	Utility	Marginal Utility
0	0	
1,000	100	100
2,000	180	80
3,000	240	60
4,000	285	45
5,000	320	35
6,000	350	30

income is distributed more uniformly between the three families, total utility rises. In case III, where income distribution is exactly equal, utility is 540, representing the maximum total utility.

This result can be derived in another way. Suppose we consider transferring \$1,000 from family 1 to family 2. The gain in utility is the marginal utility of consumption for family 2, MU_2. The loss is the marginal utility of consumption for family 1, MU_1. If family 1 is at a higher consumption level than family 2 and if marginal utility is a decreasing function of consumption, then

$$\text{Gain} = MU_2 > MU_1 = \text{loss}$$

that is, the gain exceeds the loss. Given identical individual consumption utility functions and diminishing marginal utility, whenever the distribution of consumption is unequal a redistribution toward greater equality will always result in gains exceeding losses and an improvement in social welfare. Assuming that all income is either consumed now or saved for consumption at some future date, this implies that income equality is a desirable social goal.

TABLE 2–2
Total Utility Associated with Alternative Distributions of Income

	Family	Income (Dollars)	Utility
Case I:			
	1	6,000	350
	2	0	0
	3	0	0
Total		6,000	350
Case II:			
	1	3,000	240
	2	2,000	180
	3	1,000	100
Total		6,000	520
Case III:			
	1	2,000	180
	2	2,000	180
	3	2,000	180
Total		6,000	540

The Pareto Optimum

Some economists of the neoclassical school were skeptical of the utility approach. The classical welfare criterion depended on the interpersonal comparability of the psychic gratification individuals receive from consuming goods. Yet not only is utility impossible to measure, so that the assumption that individuals have identical consumption utility functions cannot be empirically verified, but the latter proposition did not seem reasonable, even on an a priori basis. Thus the neoclassical economists felt that the utility concept was a weak reed upon which to rest welfare statements about resource allocation.

Vilfredo Pareto developed a welfare criterion that requires no comparison of individual utilities: *A* **Pareto optimum** *is a situation in which no one can be made better off without making someone else worse off.* Stated somewhat differently, if some people prefer a change and no one objects, then the change must be an improvement over the present state of affairs.

As we shall see, the Pareto criterion does not require the use of the utility concept. However, its usefulness is limited, since there are many alternatives that cannot be ranked by this criterion. Consider the familiar case of a highway authority attempting to construct a long-needed road between two growing towns. A single family refuses to sell its property, which is located directly in the path of the planned highway. The property is assessed at $25,000, and the family agrees this is a fair market value. Rerouting the highway would cost the government an estimated $500,000. The government offers the family $50,000 for the property and still it refuses to sell.

Can welfare economics provide criteria by which to judge the desirability of a court-ordered sale of the property to the government? The Pareto criterion implies that a change will not necessarily be an improvement if it harms someone. Clearly, since the family finds the change undesirable, it cannot be ranked an improvement by the Pareto criterion. On the other hand, classical welfare theory would suggest that the change would be an improvement, since it will make many people better off and only a few worse off. However, such a judgment requires comparison of the utility of the many associated with the use of the new road and the loss of utility (disutility) to the few associated with the loss of their home. The new welfare economics does not make such comparisons.

It is important not to be misled by the term *Pareto optimum,* since a Pareto optimum is not a unique allocation of resources. There are many different Pareto optimal resource allocations consistent with different distributions of utility among the individuals of a society. Suppose the government has $1 million to spend on poverty areas in the United States. One possibility is to concentrate the spending in a single area, the Appalachian region, for instance, in the hopes of making a significant impact on growth. Another is to distribute the funds uniformly among the designated poverty areas of the country. In the latter case, the impact on any one region will be less but more areas will be affected. Presumably, either program could be designed to be Pareto optimal, that is, no

resources would be wasted and no further improvements could be made without making someone worse off. Yet the distributive impact of the two programs would be quite different. Furthermore, changing from one to the other, once one is established, would not meet the Pareto criterion for an improvement, since the gain to some region(s) would be at the expense of others.[2]

Movement to a Pareto optimal resource allocation from one that is not Pareto optimal is not necessarily an improvement if a change in income distribution is involved. Suppose the Appalachian program is designed by efficiency experts with a view to eliminating all conceivable areas of waste. The alternative scheme of uniform distribution of funds is to be planned and administered locally, with a likelihood of considerable waste and inefficiency. Assuming the first program is Pareto optimal and the second is not, it would not be an improvement by the Pareto criterion to move from the second (nonoptimal program) to the first (optimal program), since some regions will lose as the Appalachian program gains. On the other hand, the Pareto criterion can be used to examine the way resources are being allocated in the second program. If the situation is nonoptimal, a better, Pareto optimal allocation can be found. Elimination of waste and inefficient techniques in the second program could make some people better off without hurting others, given the decision to distribute poverty funds on a uniform regional basis.

Contemporary Welfare Criteria

Recognizing the limitations of the Pareto criterion for ranking many alternatives, some contemporary economists have tried to develop more restrictive welfare criteria. In the spirit of the Paretian welfare economics, these criteria seek to avoid the use of the utility concept.[3]

One example of such formulation follows. Suppose, in the earlier example, the highway authority finally induces the family to sell its property by offering $100,000. Even though the family is unhappy with the move, the offer of $100,000 allows them to compensate for their loss by purchasing other goods and services. The government is better off, since it saves $400,000 of the cost of relocating the road. If the family accepts the payment voluntarily, then clearly the family prefers that option and the situation can be ranked as an improvement by the Pareto criterion. If, on the other hand, it is coerced into the arrangement, the family experiences a loss in utility and consequently interpersonal utility comparison must be made to rank the change an improvement.

[2] The term *Pareto optimum* is an unfortunate choice. We will apply the term *Pareto efficiency* to an allocation that satisfies the Pareto criterion.

[3] Attempts at developing such criteria include J. R. Hicks, "The Foundations of Welfare Economics," *Economic Journal, LXIX* (December 1939), pp. 696–712; Nicholas Kaldor, "Welfare Propositions in Economics and Interpersonal Comparisons of Utility," *Economic Journal, LXIX* (September 1939), pp. 549–552; and Tibor Scitovsky, "A Note on Welfare Propositions in Economics," *Review of Economic Studies, IX* (1941–42), pp. 77–78.

Another type of criterion explicitly introduces value judgments concerning the desirability of alternative utility (or income) distributions.[4] It might suggest that the family be forced to sell its property for $25,000 on the grounds that the social need for the highway is more urgent than the interest of the individual family. Such a criterion, however, is subjective and requires some mechanism for making the value judgment. A different arbiter might decide that the maintenance of individual freedom is worth more than the $500,000 required to reroute the road.

Several procedures have been suggested for making such value judgments, including various voting arrangements. However, a voting arrangement also requires value judgments about how votes should be distributed. For instance, if everyone gets one vote, does this violate the utility principle if income-utility relations are different for different individuals?

We will consider these contemporary welfare criteria in more detail in the final chapter of this book. In the rest of this chapter we will explore the implications of the Pareto criterion in more detail.

2–3 CONDITIONS FOR PARETO OPTIMALITY

The Pareto criterion, that a change is an improvement if someone is made better off without making anyone worse off, can be used to derive certain conditions to determine if the criterion is satisfied in a particular case. When the Pareto criterion is satisfied, resource allocation is said to be efficient.[5] Consequently, these conditions are sometimes called **efficiency conditions.**

Efficiency in Production

Pareto efficiency implies that when a certain quantity of resources is devoted to the production of a single commodity, it is always best to produce as much as possible. More is always preferred to less when no additional resources are required to produce it and when the production of other goods and services is not affected.

Consider two firms producing a single commodity, bearings. For each firm, a production function is shown in Table 2–3. A **production function** is a relation between factor input and product output. In this case, labor is the only input in the production process.[6] Technologically, firm B is clearly the more efficient producer of bearings. For any given quantity of labor input, firm B produces

---

[4] For an example of this type of analysis see Paul A. Samuelson, "Social Indifference Curves," *Quarterly Journal of Economics, LXX* (February 1956), pp. 1–22.

[5] Paretian efficiency, also called allocative efficiency, should not be confused with technological efficiency, which refers to the least cost use of inputs in a production process. While allocative efficiency implies the achievement of technological efficiency, the converse is not the case.

[6] Other inputs may be used, but we have assumed they are present in fixed amounts.

TABLE 2–3
Production of Bearings as a Function of Labor Input

FIRM A

Labor Input (100 Man-Hours)	Output (1,000 Units)	Marginal Product (MP_{LA})
0	0	
1	50	50
2	90	40
3	120	30
4	145	25
5	165	20

FIRM B

Labor Input (100 Man-Hours)	Output (1,000 Units)	Marginal Product (MP_{LB})
0	0	
1	100	100
2	175	75
3	225	50
4	255	30
5	275	20

more bearings than firm A.[7] Suppose the available supply of labor is 400 man-hours. What is the best way to allocate labor between these firms?

At first glance, it may appear that since firm B is more efficient, all the labor should be used in firm B to produce 255,000 bearings. But suppose we consider reallocating 100 man-hours to firm A. Output in firm B will fall to 225,000 and output in firm A will be 50,000. Total output rises from 255,000 to 275,000.

How do we know when a reallocation results in an improvement? One way to look at the problem is to examine the gains and losses associated with a change. The change in output associated with a stipulated change in labor input is called the **marginal product of labor** (MP_L).

In Table 2–3 marginal product is given per 100 man-hours of labor input. The gains and losses associated with a reallocation of 100 man-hours of labor between firms is measured by the marginal product. Thus, in this case, since the marginal product of labor in firm A is greater than in firm B, the gain associated with moving 100 man-hours of labor from firm B to firm A exceeds the loss, that is,

$$\text{Gain} = MP_{LA} = 50 > 30 = MP_{LB} = \text{loss}$$

[7] Superior technological efficiency may be due to superior management or more modern capital equipment. We will deal with the determinants of production relations in chapter 6.

where MP_{LA} is the marginal product of labor in A and MP_{LB} is the marginal product of labor in B.

Unless the marginal product of labor is equal in both firms, there will be some reallocation of labor for which the gain will exceed the loss or for which the loss will exceed the gain. Thus labor should be allocated between the two firms until the gain in output in the firm receiving labor is exactly equal to the loss in output in the other firm. This implies that the marginal product of labor be equal in the two firms, that is,

$$MP_{LA} = MP_{LB}$$

More generally, the marginal product of a factor input must be equal in all firms producing the same product.

If 400 man-hours are available, the most efficient allocation of resources is 300 man-hours to firm B and 100 man-hours to firm A. At this point, both would be producing where $MP_L = 50$, that is,

$$MP_{LA} = MP_{LB} = 50$$

Suppose there is a central planning board (CPB) responsible for ensuring efficient resource allocation in this economy. How could the CPB induce firms to hire the efficient amount of labor? It could set the price of labor at its marginal product. That is, each firm would be required to pay to the CPB the value of 50,000 bearings for each 100 man-hours it uses. Firm A would break even hiring 100 man-hours of labor. Firm B would hire exactly 300 man-hours. More workers would add less to output than their wage, so hiring them would incur a loss, while fewer workers would generate excess profits, and hence an increment would represent a gain. Thus the CPB could use a pricing scheme to achieve efficient resource allocation.

Suppose labor supply in the bearing industry rises to 700 man-hours. Now 100 additional man-hours should go to firm B and 200 additional man-hours to firm A, equalizing MP_L at 30. If the CPB sets the price of labor at the value of 30,000 bearings per 100 man-hours of labor and firms maximize their net gains, or profits, this allocation will be achieved.

Diminishing Returns and Efficiency

The production functions in Table 2–3 both display constantly diminishing marginal labor productivity. There are a number of reasons why we expect this **principle of diminishing returns** to apply in production. If other factor inputs are fixed, it is likely that the rate of increase of total product will slow down as more labor is applied unless the other factor inputs are also increased. It is much easier to study alone, for instance, than in a crowded library. Agricultural output increases more rapidly if fertilizer is added to newly seeded land.

Diminishing marginal productivity ensures that the efficiency condition

$$MP_{LA} = MP_{LB}$$

will result in a maximum rather than a minimum level of output. If marginal product were increasing, it would often be better to assign all labor to a single firm.[8] However, for reasons that we will consider in chapter 6, diminishing marginal productivity is likely to apply in most cases.

Alternative Production Techniques

In the previous example we were interested in maximizing the level of output from a single, limited resource. We may also be interested in the optimal utilization of *alternative* resources at each level of output. A firm generally has a number of different ways in which various resources can be combined to produce a certain level of output.

Table 2–4 shows alternative production technologies for two firms producing bearings. Suppose the total amount of capital available is 120 and the amount of labor, 180. Firm A, which is relatively more efficient in capital-intensive methods, is producing 100,000 bearings by its most capital-intensive process, using 100 capital and 20 labor. Firm B, which is relatively more efficient in labor-intensive methods, uses 20 capital and 160 labor (its most labor-intensive process) to produce 100,000 bearings. Now suppose we consider a reallocation of resources between the two firms.

Suppose firm B were to release 60 units of labor to firm A. In order to maintain production at the 100,000 level, B would require 20 units of capital. On the other hand, if A acquires 60 labor from B, it could release 60 capital and maintain production. Thus 40 units of capital are freed for use in some other productive activity. Clearly, since production can be maintained using fewer resources, the change represents an improvement. Now suppose firm B considers a move to a more capital-intensive process. Moving from B_4 to B_3 releases 40 labor and requires 20 additional capital. If the 40 released labor were used in firm A, 20 capital would be released. In other words, production will be exactly maintained at previous levels and there is no gain from a reallocation. If firm B were to move to a still more capital-intensive process, it would release 1 unit of labor for every 3 units of capital it requires. On the other hand, firm A would require 5 units of labor for every 4 units of capital released to B to maintain bearing production at the 100,000 level. Consequently, the move would result in a decline in production and would represent an inferior reallocation.

The rate at which one resource is released when another is added to produce a

---

[8] This firm need not be the firm with increasing *MP*, however. Consider the following:

	FIRM C			FIRM D	
Input	Output	MP	Input	Output	MP
1	100	100	1	25	25
2	175	75	2	75	50
3	230	55	3	130	55

If 3,000 man-hours are available, they should all be allocated to firm C for maximum efficiency, even though firm D experiences increasing marginal labor productivity.

TABLE 2–4
Alternative Resource Combinations Required to Produce 100,000 Bearings

	FIRM A		
	Capital (K_A)	Labor (L_A)	Marginal Rate of Substitution (MRS)
1.	100	20	
			$\frac{1}{2}$
2.	80	30	
			$\frac{5}{4}$
3.	40	80	
			2
4.	20	120	
			8
5.	10	200	

	FIRM B		
	Capital (K_B)	Labor (L_B)	Marginal Rate of Substitution (MRS)
1.	200	20	
			$\frac{1}{4}$
2.	120	40	
			$\frac{1}{3}$
3.	60	60	
			2
4.	40	100	
			3
5.	20	160	

fixed level of output is called the **marginal rate of factor substitution** (MRS). Suppose the marginal rate of substitution of labor for capital in firm B is greater than that for firm A, that is,

$$MRS^B_{LK} > MRS^A_{LK}$$

If B moves to a more capital-intensive process, more labor is released from B than is required by A to maintain production at the reduced level of capital. If B moves to a more labor-intensive process, it will require more labor than can be released from A if A is to maintain production. Only in the case where the marginal rate of factor substitution is equal for the two firms, that is,

$$MRS^B_{LK} = MRS^A_{LK}$$

will the amount of one factor released from one firm be the exact amount required to maintain production in the other. Consequently, no reallocation of factors would represent an improvement. Equal marginal rates of factor substitution for all firms producing the same product is the second Pareto efficiency criterion.

Notice that to derive this result we assumed that the MRS_{LK} is constantly increasing as processes become more labor-intensive. This implies that the more labor is being used relative to capital the more labor can be released as new capital is acquired without affecting the level of output. It is consistent with the principle of diminishing returns that the more abundant one factor is relative to another the less will one more unit add to total output. The more fertilizer we have applied to an acre of land the smaller will be the impact on output of an additional application.

How could the central planning board ensure that all firms will produce at equal marginal rates of factor substitution? In this case, the CPB could announce that all firms producing bearings may acquire all the capital and labor resources they wish but that the price of capital is twice that of labor. Firms would acquire resources until no further gains or losses are realized. This would occur where the acquisition of an additional unit of capital is equivalent to a loss of two units of labor. By setting the price of capital at twice that of labor, the CPB can induce firms to adopt the production technology for which a unit of capital is worth two units of labor. The second efficiency criterion,

$$MRS^B_{LK} = MRS^A_{LK}$$

established that this is the optimal production plan for the firms in this problem.

Alternative Production Possibilities

Suppose we consider firms that are producing more than one good: Two firms are producing bearings and screws, using a single resource, labor. Table 2–5 shows the alternative output combinations that each firm can produce per man-hour of labor input.

Imagine that firm A is producing 40 bearings and 45 screws per man-hour and firm B is producing 30 bearings and 27 screws per man-hour, and that we desire a reallocation of resources to increase screw production. In firm A, when the production of screws is increased by 10, bearing output falls by 20. For this firm, the cost of producing 10 additional screws is the forgone revenue from the sale of 20 bearings. Put in another way, the cost of screws in terms of bearings is 2. When 10 additional screws are produced in firm B, bearing output falls by only 10. In B the bearing cost per screw is 1. Given the initial levels of output, firm B is the low cost producer of screws, cost being measured in terms of foregone bearing production. The bearing cost of screws is called the **marginal rate of transformation** (*MRT*). Clearly, if an increase in screw output is desired, the increase should come from the firm with the lowest *MRT*.

Suppose no increase in screw production is desired. Is there any reallocation

TABLE 2–5
Alternative Combinations of Bearings and Screws that Can Be Produced Per Man-Hour of Labor

FIRM A

Bearings	Screws	Marginal Rate of Transformation $\left(-\frac{\Delta B}{\Delta S}\right)$
100	0	
		1
80	20	
		$1\frac{1}{3}$
60	35	
		2
40	45	
		4
20	50	
		10
0	52	

FIRM B

Bearings	Screws	Marginal Rate of Transformation $\left(-\frac{\Delta B}{\Delta S}\right)$
50	0	
		$\frac{2}{3}$
40	15	
		$\frac{5}{6}$
30	27	
		1
20	37	
		$1\frac{1}{4}$
10	45	
		$2\frac{1}{4}$
0	51	

of resources within firms that could increase the output of screws using no additional resources? Firm B can produce screws at a lower bearing cost than A. On the other hand, the screw cost of bearings is lower for A. If B were to produce 10 more screws and A 10 fewer screws, bearing output would fall by 10 in B and rise by 20 in A. This would result in a net increase of 10 bearings with no decline in screw production. Although each firm must give up screws to produce additional

bearings, a reallocation of resources among several firms can increase bearing output with no sacrifice of screws. This will be the case as long as *MRT*, that is, the bearing price of screws, is unequal among the firms.

Suppose that the bearing cost of screws is greater in firm A than in firm B, that is,

$$MRT_A = \left(-\frac{\Delta B}{\Delta S}\right)_A > \left(-\frac{\Delta B}{\Delta S}\right)_B = MRT_B$$

where ΔB refers to a change in bearing production and ΔS refers to a change in screw production. If resources are reallocated so that net screw production is left unchanged, that is,

$$-\Delta S_A = \Delta S_B$$

then bearing production in firm A will increase by more than it decreases in B, that is,

$$\Delta B_A > -\Delta B_B$$

As long as *MRT* is different among firms, a reallocation of resources within firms will always result in an improvement. Thus the third efficiency condition can be restated

$$MRT_A = MRT_B$$

Another way to state this criterion is that the opportunity cost of one good in terms of another must be the same for all firms producing the goods. Opportunity cost refers to the output of a good forgone when another is produced.

The statement that all firms should be producing at equal *MRT* is another form of the well-known **principle of comparative advantage** in international trade. Notice that firm B is a relatively inefficient producer of both screws and bearings, that is, output per man-hour is lower than in firm A. Yet in terms of bearings, B can produce screws cheaper than A. If A and B were two countries trading, it would pay B to produce more screws and obtain its bearings from A. B could obtain bearings from A at a price of one-half screw per bearing,[9] while it would cost one screw per bearing to produce bearings domestically.

If certain resources or resource combinations were better suited to bearing than to screw production, it is likely that the bearing cost of screws would rise as more screws are produced. If *MRT* rose in both firms (or countries), it is probable that the optimal degree of specialization would require both to produce some of each commodity, although one firm (or country) might produce relatively more of one commodity than the other. On the other hand, if *MRT* were constant for both firms or countries (as is the case in most examples from international trade textbooks), then optimal specialization would entail complete specialization

[9] The actual price of bearings will lie between one-half screw and one screw, depending upon the elasticity of reciprocal demand for screws and bearings in the two countries.

in the production of that commodity for which MRT is lowest. Notice that equality of MRTs will only occur when MRT is not constant, and the criterion results in an optimum only when MRT rises as more of a good is being produced. If the opportunity cost of producing bearings were to fall the more bearings were produced, then complete specialization would be the most desirable way to organize production. As we will see in chapter 6, this is unlikely to be the case.

If the central planning board were to establish a fixed bearing price for screws, then firms would reorganize production until

$$MRT = p_S$$

where p_S is the bearing price of screws. If MRT were less than p_S, firms would increase screw output, since the addition to revenue, p_S, would be greater than the addition to cost, MRT. On the other hand, firms would cut back screw production if the bearing cost (MRT) exceeded the price. Since MRT rises and falls with screw production, adjustments would continue until MRT equaled p_S. If all firms could sell screws at p_S, all firms would be producing at equal marginal rates of transformation. Thus, the CPB could establish efficiency in production.

Efficiency in Consumption

The Pareto criterion states that a change will be an improvement if someone can be made better off without making someone else worse off. In order to establish welfare changes associated with production alternatives, we must consider how such changes are evaluated by households.

Consider two neighbors, Tom and Jerry, who are do-it-yourself repairmen, maintaining home inventories of hardware, including screws and bearings. Tom proposes a trade of screws for bearings. He is willing to trade 3 bearings for 1 screw. Jerry, on the other hand, would be willing to accept 2 bearings for 1 screw. Suppose Tom acquires 10 screws from Jerry for 30 bearings. Tom is no worse off, or he would not have made the trade. On the other hand, Jerry is better off, since he would have made the trade for 20 bearings and he actually received 30, a bonus of 10.

The terms on which an individual is willing to trade one good for another is called the **marginal rate of product substitution** (MRS). Clearly, at least one party will benefit from a trade if the MRS's of the two individuals differ. If I am twice as fond of apples as of oranges and you like apples and oranges equally well, then, if we both have apples and oranges, we could both gain from a trade. You will be better off if I give you more than one orange per apple, and I will be better off giving you anything less than two oranges per apple. The terms on which the trade is actually made will fall somewhere between the different marginal rates of product substitution. If these rates are the same, then the exchange would be made at MRS, which would leave everyone exactly as well off as before the trade. Equality of the marginal rate of product substitution between any two goods for all individuals in the economy, that is,

$$MRS^{\mathrm{I}} = MRS^{\mathrm{II}}$$

is the fourth Pareto criterion. It implies that no further trade or reallocation can be made to make someone better off without making someone else worse off.

In this example, the terms of trade for bearings and screws were established by the trading partners. Suppose a bearing price for screws were set by the central planning board. Persons whose marginal rates of substitution were greater than the set price would buy screws and those with lower *MRS* would sell them. Clearly *MRS*, that is, the relative preference for screws over bearings, will decrease the more screws and the fewer bearings an individual possesses. As households with high *MRS* acquire more screws and those with low *MRS* more bearings, their relative preferences at the margin move to equality. Although the consumption pattern of some households may be relatively "screw-intensive" and that of others more "bearing-intensive," at the margin (for the final trade) each household's *MRS* will exactly equal the terms of trade for all households that consume some of both goods. Thus the CPB, by offering to buy and sell screws at a fixed bearing price, can ensure efficiency in consumption.

Global Efficiency

So far we have defined the conditions for Pareto optimal resource utilization in production and for Pareto optimal distribution in consumption.[10] We still must determine that what is produced satisfies the requirements of consumers.

Recall that the marginal rate of transformation between two goods (which must be equal for all firms producing both goods if efficiency is to be achieved) tells us how much of one we must give up to obtain one more unit of the other. On the other hand, the marginal rate of product substitution (which must be equal for all households consuming both goods if efficiency is to be achieved) tells us how much of one good the household is willing to forgo in order to obtain the other without feeling worse off.

The fifth Pareto criterion states that the optimal scale of production of the two goods from the point of view of the consuming sector occurs where the marginal rate of product substitution equals the marginal rate of transformation, that is,

$$MRS = MRT$$

To see why this must be so, consider the possibility that the marginal rate of substitution is greater than the marginal rate of transformation, that is,

$$-\frac{\Delta B^H}{\Delta S^H} = MRS > MRT = -\frac{\Delta B^F}{\Delta S^F}$$

This implies that households are willing to give up more of one good for the other than is necessary given the production alternatives. Consequently, they would

[10] Note that Pareto optimal distribution in consumption can be defined for any distribution of income. This refers only to the distribution of particular goods and services among individuals given the distribution of purchasing power, or income.

gain if more resources were allocated to produce that good, say bearings. As bearings production rises, so does *MRT*, while *MRS* falls as households acquire more bearings. On the other hand, if

$$MRS < MRT$$

the cost of producing more bearings is greater than households are willing to pay. Fewer bearings (and more screws) should be produced until the marginal rates of substitution and transformation are brought into equality.[11]

Another way to look at the question of global efficiency is to evaluate the gains and losses associated with a reallocation. Suppose *MRS* of bearings for screws is 3, while *MRT* is 2, that is,

$$-\frac{\Delta B^H}{\Delta S^H} = MRS = 3 > 2 = MRT = -\frac{\Delta B^F}{\Delta S^F}$$

Households are willing to give up 3 bearings to obtain 1 screw. This means the welfare value of the screw is 3 bearings. The cost of producing a screw is 2 bearings. Thus, if resources are reallocated from bearing to screw production, the gain (bearing value) is 3, while the loss is 2. The gain, or *MRS*, exceeds the loss, or *MRT*, and so the change will be an improvement.

Global Efficiency and Absolute Levels of Production

Suppose the central planning board agrees to buy and sell screws at a fixed price, p_S. Producers will be induced to reallocate resources within firms until

$$MRT = p_S$$

Households will buy and sell screws for bearings until

$$MRS = p_S$$

As long as the CPB sets the same price for both firms and households, and the former argument applies where the reallocation is to more screws and fewer bearings, this will ensure that

$$MRS = p_S = MRT$$

and the condition for global efficiency would appear to be satisfied.

However, the establishment of fixed prices at which commodities can be

---

[11] Note that if

$$-\frac{\Delta B^H}{\Delta S^H} = MRS_{BS} < MRT_{BS} = -\frac{\Delta B^F}{\Delta S^F}$$

where MRS_{BS} is the marginal rate of substitution of bearings for screws, then

$$-\frac{\Delta S^H}{\Delta B^H} = MRS_{SB} > MRT_{SB} = -\frac{\Delta S^F}{\Delta B^F}$$

where MRS_{SB} is the marginal rate of substitution of screws for bearings.

bought and sold only ensures that at the margin, *changes* in screw and bearing output will be matched by *changes* in consumer purchases, that is,

$$-\frac{\Delta B^H}{\Delta S^H} = MRS = p_S = MRT = -\frac{\Delta B^F}{\Delta S^F}$$

It does not ensure that the actual *level* of screw and bearing production will be equal to the total amount consumers wish to purchase at p_S, that is,

$$S^H = S^F \qquad \text{and} \qquad B^H = B^F$$

But global efficiency does require that the amount consumers wish to purchase at p_S be exactly equal to the amount produced at p_S. If consumers are not willing to purchase all the screws produced at p_S, then there will be an excess supply of screws, which is wasteful in the face of scarcity. Fewer resources should have been allocated to screws to begin with.

Suppose

$$MRS = 2.5 = MRT$$

and the CPB sets

$$p_S = 2.5$$

When MRT is 2.5, firms produce 20,000 bearings and 50,000 screws. MRS is 2.5 when consumers purchase 45,000 bearings and 40,000 screws. Thus, when p_S is 2.5, 20,000 bearings and 50,000 screws will be produced. But households will only purchase 40,000 screws, leaving an excess supply of 10,000. Furthermore, households will attempt to purchase 45,000 bearings when only 20,000 are available. Thus the establishment of fixed commodity prices by the CPB does not necessarily produce global efficiency. For global efficiency the condition

$$MRS = MRT$$

must be met, and *in addition*, the amount of each good produced must equal the amount consumers wish to purchase, that is,

$$S^H = S^F \qquad \text{and} \qquad B^H = B^F$$

We will return to this question in chapters 3 and 4.

Optimal Supply of Resources

The principle of equalizing marginal rates of product substitution among goods with their marginal rates of transformation for global efficiency can also be applied to the theory of optimal factor supply. Earlier we saw that although resources are scarce, they nevertheless may be used for other than productive purposes in ways that contribute to social welfare. An alternative to the use of labor is leisure. Resources devoted to the production of capital goods can be used for the more immediate production of consumer goods.

Consider the optimal amount of labor force participation. From a different

perspective, this is equivalent to the optimal consumption of leisure. A household has a certain amount of the labor resource available. Conceivably both husband and wife could work twenty-four hours each. However, with maximum work effort the welfare value of leisure would be nearly infinite, since such efforts would result in rapid physical deterioration. Given the number of available hours, the household must decide whether to allocate them to labor force participation, for which it receives income, or to leisure, which provides welfare value. Since income can be used to purchase goods, both income and leisure provide welfare value. The marginal rate of substitution of income for leisure is the rate at which a household is willing to sacrifice income for leisure. For instance, a family may be willing to forgo $3 of income for an additional hour of leisure. Furthermore, *MRS* of income for leisure (MRS_{Yl}) is likely to increase the more hours are devoted to work and the fewer to leisure.

Our Pareto criterion suggests that optimal consumption of leisure occurs where the marginal rate of substitution of income for leisure equals the opportunity cost of leisure in terms of income, that is,

$$MRS_{Yl} = MRT_{Yl}$$

The marginal rate of transformation of income for leisure is the amount of income (in the form of goods and services) forgone when one man-hour is withheld from the labor force. From the viewpoint of the whole economy the amount of goods and services forgone when labor is reduced is the marginal product of labor, MP_L. Thus the efficiency condition for optimal labor supply can be expressed as

$$MRS_{Yl} = MP_L$$

Furthermore, the demand for labor must equal the supply when the marginal condition is met.

Suppose the central planning board sets a fixed wage at which firms hire labor. From the viewpoint of the households the wage represents the price they must pay to consume leisure. They will consume leisure until the welfare value of the last hour, MRS_{Yl}, exactly equals its price, that is, until

$$MRS_{Yl} = w$$

Firms will hire labor until the opportunity value of the last worker hired in terms of output, MP_L, equals the wage, that is, until

$$MP_L = w$$

Thus the establishment of a fixed wage ensures that the marginal condition for the optimal degree of labor force participation is met, since the condition

$$MRS_{Yl} = w = MP_L$$

is satisfied. However, setting a fixed wage does not ensure that the demand for labor will equal the supply. There may be involuntary unemployment of labor

(excess supply) even though the marginal condition for global efficiency is satisfied.

Allocation of Resources to Investment

A similar analysis can be made for the efficient allocation of resources to producing capital equipment. Capital may consist of physical plant and machines that are used to produce other goods and services. Another way to devote resources to indirect production of goods is by training labor to perform more productively. Investment of current labor resources for making labor more productive is called investment in **human capital.** One way to explain why different people earn different rates of pay is that they have acquired different amounts of human capital. Thus part of their income is not a pure wage but is a return on human capital.

Suppose a household wishes to invest part of its available labor-leisure time in the acquisition of human capital. The household may want to decide whether the mother should go to college or whether she should go to work as a secretary. Assuming that she is indifferent to the way in which she uses her time (that is, she has no preference for secretarial work per se over studying), the monetary gain to the family of investment in human capital (education) is the difference in income she could expect as a college graduate over her secretarial pay. However, to obtain the increase in income the family must wait several years. Most people have a subjective preference for current consumption over future consumption. The rate at which a household is willing to exchange future consumption for current consumption is called the marginal rate of substitution of future for present consumption, MRS_{fp}. The MRS_{fp} is generally greater than 1, that is, a household will give up $1 of consumption today only if it can expect to receive more than $1 tomorrow.[12] Thus

$$MRS_{fp} = 1 + \tau$$

where τ is the extra amount of future consumption required to compensate the household for giving up current consumption.

The Pareto criterion suggests that the efficient level of capital accumulation occurs where the marginal rate of substitution of future for present consumption equals the marginal rate of transformation, that is,

$$MRS_{fp} = MRT_{fp}$$

The rate at which future consumption can be transformed into present consumption, MRT_{fp}, is $1 + \rho$, where ρ is the rate of return on capital. For human capital, this is the per cent increase in expected income as a result of training.

[12] This helps to explain why borrowers must pay interest to lenders.

For physical capital, this is sometimes called the marginal productivity of capital, MP_K.[13] Thus efficiency occurs where

$$MRS_{fp} = 1 + \tau = 1 + \rho = MRT_{fp}$$

The family should invest in a college education for the mother as long as the rate at which it is willing to exchange future for present consumption is greater than the expected increase in her income after graduation. Alternatively, investment should take place where

$$\tau = \rho$$

For global efficiency in the allocation of resources to capital goods, not only must the marginal condition be met but the amount of resources households wish to release from current consumption must be equal to the amount firms wish to invest. The central planning board could influence firms' and households' decisions to use current resources for capital accumulation by setting a fixed rate of interest i. The quantity $1 + i$ represents the opportunity cost to households of current spending. Thus households will borrow or lend until the subjective rate at which they would exchange the last dollar of future consumption for present consumption is equal to the opportunity cost of present consumption, that is,

$$1 + \tau = MRS_{fp} = 1 + i$$

or

$$\tau = i$$

The quantity $1 + i$ is also the opportunity cost for firms investing current resources, since these same resources could be sold and the proceeds loaned at interest. Thus capital accumulation will take place until the return from the last unit is equal to the interest rate, that is,

$$\rho = i$$

or

$$1 + \rho = MRT_{fp} = 1 + i$$

The establishment of a fixed rate of interest by the CPB thus ensures that the marginal condition for the efficient rate of capital accumulation is met, since the condition

$$MRS_{fp} = 1 + i = MRT_{fp}$$

is satisfied. However, setting a fixed rate of interest does not ensure that the supply of resources from current consumption will be equal to the demand at that interest rate. Some mechanism other than a pricing arrangement must be introduced to ensure that supply equals demand in product and factor markets.

[13] Alternatively, the compound rate of return on capital investment is sometimes called the internal rate of return or the marginal efficiency of capital.

2–4 THE MARGINAL CONDITIONS AND ECONOMIC EFFICIENCY

It is important to keep in mind that the so-called marginal conditions for economic efficiency are derived from the Pareto criterion:

A change is an improvement if someone is made better off without making anyone else worse off.

The marginal conditions are merely rules of thumb for achieving the Pareto optimum. They are only valid as rules of thumb if certain assumptions are met. For instance, if opportunity costs do not increase as the output of a good increases, the condition that all firms producing the good operate at the same marginal rate of transformation produces a minimum rather than maximum level of output. Furthermore, if all costs of production are not accounted for in calculating the firm's marginal rate of transformation, then the marginal condition will not produce optimal results. This may occur if the smoke from a factory inflicts a real or psychic cost on nearby residents but is not included in calculating the production cost of the good by the firm managers.

Thus, while the Pareto criterion remains valid for making certain types of welfare comparisons,[14] the marginal conditions only reflect Pareto optimality under certain circumstances. Since the requisite assumptions are satisfied in many cases, economists often treat the marginal conditions as welfare criteria in themselves. However, it must be kept in mind that the attainment of the marginal conditions is only a means (in some cases) of fulfilling a larger objective, not an end in itself. In chapter 3 we will consider cases in which attainment of the marginal conditions does not imply efficient resource allocation.

QUESTIONS FOR STUDY AND REVIEW

1. Two firms use labor to produce widgets. Their production functions are as follows:

FIRM A		FIRM B	
Labor Input	Widgets	Labor Input	Widgets
1	500	1	300
2	900	2	550
3	1,250	3	700
4	1,500	4	800
5	1,600	5	850
6	1,675	6	875

a) Which firm is the low cost producer?
b) If 6 units of labor are available, how should they be distributed between the firms to ensure efficient use of resources?
c) What price should the central planning board charge the firms for the use of labor?
d) If the labor supply increases to 9, what price should the central planning board

[14] Remember that a Pareto optimum may be associated with an undesirable distribution of income.

charge the firms for the use of labor to ensure efficient resource allocation in the widget industry?

e) Suppose technological change in the widget industry reduces by one half the labor input requirement for each level of production for both firms. How would your answers to the above questions be affected?

2. "American agriculture is the most efficient in the world." Assuming that this is true, would the United States ever import agricultural goods? Be sure to justify your answer.
3. If economics is concerned with finding ways to achieve maximum *total* satisfaction from its limited resources, why are economists so concerned with *marginal* relations?
4. The idea that economic efficiency should be the most important *economic* goal of any society has been defended and denied. Present your views on the controversy, with some discussion on both sides of the question.
5. The principle of comparative advantage is usually associated with the theory of international trade. Generalize the principle and show its relation to the theory of resource allocation.
6. "If efficient resource allocation is a goal of an economy, then prices must exist in that economy." Comment.

ADDITIONAL READING

Bator, Francis M. "The Simple Analytics of Welfare Maximization," *American Economic Review, XLVII* (March 1957), pp. 22–59.

Baumol, William J. *Economic Theory and Operations Analysis,* 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1972, chapter 16.

Kohler, Heinz. *Welfare and Planning.* New York: Wiley, 1966, chapter 2.

3 Economic Efficiency in a Market System

An economic system is a set of institutions and relations to allocate resources. Thus an economic system must determine what is to be produced, how it is to be produced, and to whom the product is to be distributed. Such decisions may be explicit or implicit. A central planning board operating on the basis of certain explicit criteria may assign target output levels for all industries, stipulate the way in which productive factors are used by individual firms, and ration the final products to households. Since allocative decisions are explicit, it is possible to ascertain if the predetermined allocative criteria are satisfied.

Where allocative decisions are made implicitly, as they are in a market system, a theoretical model is needed to demonstrate that the resultant allocation of resources is optimal. This was the contribution of Adam Smith in *The Wealth of Nations.* Although in a market economy allocative decisions are not made explicitly by a central planning board, or by tradition as in the medieval society, Smith attempted to show that the invisible hand of the market would produce a socially desirable resource allocation.

The possibility of efficient resource allocation is not, however, exclusive to market systems as Adam Smith had implied. There are other ways to organize production and distribution that will produce optimal allocative decisions. Furthermore, a market system may fail in many ways to achieve Pareto optimality. In this chapter we examine a market model in terms of the efficiency criteria. In chapter 4 we will discuss alternative economic systems.

3–1 THE THEORETICAL JUSTIFICATION FOR MARKETS

Consider an economy in which all goods and services as well as factors of production are traded in markets. A market is a theoretical construct in which a single price for a good or service is established. Market participants are influenced by prices in making their sales and purchases. If the quantity offered for sale is more or less than that

desired by buyers at the prevailing market price, buyers and sellers will readjust their bids and offer prices until an **equilibrium** price is established. At such an equilibrium the exact quantity offered for sale is purchased by buyers and there is no additional demand for the good. Consequently, there will be no further readjustments in price by either buyers or sellers.

In this section we examine the way in which market participants—households and firms—decide the amount of goods and factor services to buy and sell on the basis of market prices. In particular, we are interested in whether or not their market decisions are consistent with the efficiency conditions set forth in chapter 2.

Efficiency in Product Markets

The Household

Since market prices are fixed by the interaction of many buyers and sellers, a household views these prices as given data in its decisions concerning how much of each good to purchase and the amount of factor services to supply. The purchasing power of each household is small in relation to the total demand for products, so the decision of any one household to rearrange its purchases will have a negligible effect on market price.

Households will view market prices as the opportunity costs of goods in terms of each other. If apples cost 5 cents and bananas 10 cents the opportunity cost of apples in terms of bananas is 2, the ratio of their prices, that is,

$$-\frac{\Delta A}{\Delta B} = \frac{p_B}{p_A}$$

Households will allocate their budgets between apples and bananas where they are equally happy with 2 apples and 1 banana, or equivalently where

$$MRS_{AB} = -\frac{\Delta A}{\Delta B} = \frac{p_B}{p_A}$$

Thus, between any two goods, X and Y, households will adjust their purchases until the MRS_{YX} is equal to the ratio of the product prices.

Notice that since p_X and p_Y are set in the markets for X and Y, all household units face identical prices, so that

$$MRS^{\mathrm{I}}_{YX} = \frac{p_X}{p_Y} = MRS^{\mathrm{II}}_{YX}$$

All households will adjust their purchases of X and Y until their marginal rates of product substitution are equal, satisfying the marginal condition for efficiency in consumption. This does not imply that all households purchase the same quantities of X and Y, but at the margin each household is willing to exchange additional units at the same rate.

The Firm

A similar analysis can be made of the behavior of firms. Suppose there are many sellers in the market and the output of any single firm is so small in relation to total market supply that its decision to increase or decrease output will have a negligible effect on market price. Under these conditions firms view selling price as fixed, since it is established by the interaction of many buyers and sellers in the market.

Firms in a market economy are assumed to adjust output levels to maximize their profits. The maximum profit level is reached when the firm can no longer increase output without increasing costs above the selling price of the good. The increase in cost associated with the production of an additional unit of output is called **marginal cost.** Marginal cost is generally assumed to rise with increasing output because of diminishing returns to fixed plant and equipment as additional labor and materials are employed. Marginal cost also increases if factor prices rise when the demand for those factors increases. Thus the maximum profit output occurs where price equals marginal cost, that is,

$$p = MC$$

The ratio of the marginal costs of any two goods is their marginal rate of transformation, that is,

$$-\frac{\Delta Y}{\Delta X} = \frac{MC_X}{MC_Y} = MRT_{YX}$$

Suppose the marginal cost of X is twice the marginal cost of Y. If we are willing to give up a unit of X, we can produce two additional Y at the same cost. Since profit-maximizing firms produce where marginal cost equals the price established in the market, that is,

$$p_X = MC_X \qquad \text{and} \qquad p_Y = MC_Y$$

it follows that

$$\frac{MC_X}{MC_Y} = \frac{p_X}{p_Y} = MRT_{YX}$$

Since all firms can sell their output at these prices, this implies that

$$MRT^A_{YX} = \frac{p_X}{p_Y} = MRT^B_{YX}$$

That is, X and Y will be produced at the same marginal rate of transformation throughout the economy. Thus another efficiency condition is satisfied.

Global Efficiency in Product Markets

A market system ensures that the two criteria for global efficiency are met. First, the market mechanism, by which bid and offer prices are continually readjusted

until quantity supplied equals quantity demanded, ensures that there will be no excess supply of, or demand for, goods and services when the market is in equilibrium.

Furthermore, the marginal condition is also satisfied. Since the condition

$$MRS_{YX} = \frac{p_X}{p_Y}$$

is met in the household sector and the condition

$$MRT_{YX} = \frac{p_X}{p_Y}$$

is met in the producing sector, then when buyers and sellers trade at the market prices, the condition

$$MRS_{YX} = \frac{p_X}{p_Y} = MRT_{YX}$$

is also met. The price mechanism will ensure that the rate at which consumers are willing to exchange commodities is equal to the relative costs of the last units produced.

In a market system, therefore, the marginal condition for global efficiency is satisfied, and in addition there is no excess demand for, or supply of, goods and services. Thus global efficiency is achieved.

Efficiency in Factor Markets

As long as buyers and sellers act independently and each is so small in relation to the total market that his own actions will not perceptibly alter market prices, each adjusts his behavior with established market prices as given data.

The Firm

Firms, maximizing profits, will hire factors of production as long as the addition to revenue exceeds the addition to cost. The addition to revenue when a factor is employed is its marginal product multiplied by the product price (the **marginal value product**). The cost of employing the factor is the factor price established in the market. Thus, assuming the marginal product decreases as more of the factor is employed (in accordance with the principle of diminishing returns), the profit-maximizing firm will employ more of the factor until the marginal value product equals the factor price, that is,

$$p \times MP_f = p_f$$

where p is the product price, MP_f is the marginal factor productivity, and p_f is the factor price.

Since the factor price is established in the market, all firms must pay the same price. For two firms, A and B, producing the same good,

$$p \times MP_{fA} = p_f = p \times MP_{fB}$$

This implies

$$MP_{fA} = MP_{fB}$$

that is, marginal factor productivity is the same for both firms.

If more than one factor is employed (L and K, for instance), then

$$\frac{MP_{LA}}{MP_{KA}} = \frac{p_L}{p_K} = \frac{MP_{LB}}{MP_{KB}}$$

The ratio of marginal factor productivities is equal to the marginal rate of factor substitution, the rate at which factors can be substituted in production without altering the level of output. Suppose the last unit of labor adds 10 units of output, while the last unit of capital adds 5 units. We could substitute 2 capital per unit of labor without affecting output, that is,

$$-\frac{\Delta K}{\Delta L} = MRS_{KL} = \frac{MP_L}{MP_K} = 2$$

Since all firms must purchase factors of production at the prices established in the factor markets, profit-maximizing firms producing the same product will operate at equal marginal rates of factor substitution, that is,

$$MRS^A_{KL} = \frac{p_L}{p_K} = MRS^B_{KL}$$

The Household

Households will consume leisure where the marginal rate of substitution between other goods and leisure is equal to the price of leisure in terms of other goods. Income, Y, can be viewed as generalized purchasing power for goods other than leisure. The price of leisure in terms of other goods is the amount of income forgone when more leisure is consumed, the wage. Thus, households will consume leisure where the marginal rate of substitution of income for leisure equals the wage established in the labor market, that is,

$$MRS_{Yl} = w$$

Since this condition holds for the demand for leisure, it also holds for its mirror image, the supply of labor. Since all households must accept the wage established in the labor market, they will all supply labor (consume leisure) at equal marginal rates of substitution of income for leisure, that is,

$$MRS^{\mathrm{I}}_{Yl} = w = MRS^{\mathrm{II}}_{Yl}$$

Global Efficiency in Factor Markets

Under the conditions described, the two criteria for global efficiency are met in factor markets as well as in product markets. The market mechanism ensures that factor prices will be continually readjusted until there is no excess supply of, or demand for, productive factors.

In addition, the marginal conditions are satisfied. Firms will hire labor until the marginal value product is equal to the wage, that is,

$$p \times MP_L = w$$

If all prices are expressed in dollars, then the marginal value product of labor is the dollar value of the change in output when the last worker is hired. Thus the marginal value product is the marginal dollar product, $MP_L(\$)$, that is,

$$p \times MP_L = MP_L(\$)$$

and

$$w = MP_L(\$)$$

When the wage is established in the labor market, households will supply labor until the marginal rate of substitution is equal to the wage, that is,

$$MRS_{Yl} = w$$

Since producers pay the same wage that households receive, the marginal rate of substitution of dollars (income) for leisure by households will be equated to the marginal dollar product of labor, that is,

$$MRS_{Yl} = w = MP_L(\$)$$

which is the marginal condition for global efficiency in the labor market.

Similarly, it can be shown that if an interest rate is established in the capital market that is the same for all lenders and borrowers, the marginal rate of substitution of future for present dollars by households will be equated to the marginal rate of transformation of future for present dollars by firms, since

$$MRS_{fp} = 1 + i = MRT_{fp}$$

Thus, in a market system, the marginal conditions for efficiency are satisfied in both factor markets and the markets for goods and services. The market mechanism itself ensures that there will be no excess demand for, or supply of, any good or service.

3–2 MARKET FAILURES

In the previous section we developed a market model that theoretically satisfies all the Paretian conditions for economic efficiency. Such a model is traditionally set forth to justify a market system as a means of efficient resource allocation.

Proponents of a market system and laissez faire public policy argue that the market mechanism will produce an equilibrium situation in which no one can be made better off without making someone else worse off. However, it is also well known that there are many cases in which an actual market system will not result in a welfare optimum. This is because the theoretical model contains certain implicit assumptions that are not always satisfied in the real world. In this section we examine some of these market failures.

Failure to Achieve Equilibrium

It was noted that the achievement of market equilibrium in which the quantity demanded equals the quantity supplied is necessary for global efficiency. An assumption implicit in our discussion of a market system is that market equilibrium will always be achieved within a reasonable time frame. Producers will respond to excess supply by lowering supply price, which will induce consumers to purchase more and hence clear the market. If there is excess demand, consumers will bid up prices to induce firms to increase supply. Thus prices are presumed to be flexible in any direction and producers' and consumers' responses to price changes are seen as being sufficiently rapid to clear the market within a reasonable period of time.

An example of the misapplication of this market model was the notion that the chronic unemployment of the 1930s could be remedied by following laissez faire policies. Traditional economists argued that if labor was in excess supply, workers would bid down wages, causing firms to hire more workers, thereby eliminating unemployment. Thus laissez faire would eventually result in a full employment equilibrium. John Maynard Keynes in his revolutionary treatise, *The General Theory of Employment, Interest, and Money,* pointed out that in a modern industrial economy wages tend to be inflexible in the downward direction. Furthermore, when demand is severely depressed, firms are likely to respond slowly, if at all, to lower factor prices. Keynes's famous statement, "In the long run we are dead," underscored the fact that the static analysis of markets ignores the amount of time required to achieve equilibrium.[1] A laissez faire policy might produce results in theory, but the amount of time required to achieve equilibrium is unspecified. Public policy must be addressed to eliminating unemployment immediately, not in fifty or one hundred years.

Other economists have pointed out limitations of the static analysis that assumes the economy is always in equilibrium.[2] A dynamic approach that looks at market behavior in the short run is essential for an examination of the welfare effects of a market system.

---

[1] This is an oversimplification of the Keynesian analysis of unemployment. For a fuller discussion see Nancy S. Barrett, *The Theory of Macroeconomic Policy* (Englewood Cliffs, N.J.: Prentice-Hall, 1972), chapters 8 and 9.

[2] See, for instance, Paul A. Samuelson, *Foundations of Economic Analysis* (Cambridge, Mass.: Harvard University Press, 1947), chapter 9.

The question of equilibrium analysis in economic theory is an important methodological issue that cannot be fully examined in the context of an overview of market failures. We will return to this issue in chapter 7.

Monopoly

A crucial assumption of the market model is that buyers and sellers act independently and that each is so small in relation to total market demand and supply that the purchases or sales of any one has a negligible effect on market prices.

This assumption flies in the face of the observed tendency in many markets for buyers or sellers to join together in an attempt to influence market prices. Each individual worker is small in relation to the total labor market, and his decision to work more or less has no impact on prevailing wages. Labor unions, however, can have a significant effect on market wages. By a collective decision to withhold labor in a strike or by restricting entry of workers into the market, unions can force firms to pay wages higher than would be established in the market model.

The unionization of labor in a market economy is the natural outgrowth of another phenomenon at odds with the assumptions of the model—the growth of firms to mammoth monoliths that account for such a large proportion of industry output that they can effectively control market prices by their decisions concerning how much to produce.[3]

What is the effect on the allocation of resources when a buyer or seller is large enough to influence market prices by his purchases or sales? Suppose the goal of the firm is to maximize profits. Just as in the previous analysis, the firm will expand output as long as the addition to revenue, marginal revenue (MR), exceeds the addition to cost, marginal cost, MC. If marginal cost increases with output, profit maximization occurs where

$$MR = MC$$

For the small firm whose output decisions cannot affect market price, MR is equal to price, that is,

$$MR = p$$

since the price is the change in revenue associated with selling one more unit. Such a firm can sell as much as it can produce at that price, and its demand curve, the relation between its output and potential selling price, is a horizontal line, as shown in Figure 3–1, intersecting the vertical axis at the prevailing market price, p_e. For such a firm, profit maximization occurs where price equals marginal cost, that is,

$$MR = p = MC$$

[3] John Kenneth Galbraith, *American Capitalism: The Theory of Countervailing Power* (Boston: Houghton Mifflin, 1956) discusses this tendency for units on one side of a market to join together in response to monopoly power on the other side.

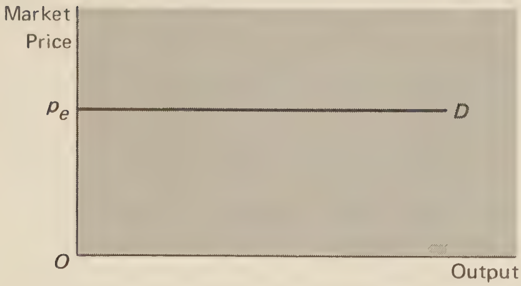


FIGURE 3–1
Relation Between Market Price and Output for a Small Firm (Market Model)

Suppose, however, that the firm is so large that it can effect changes in market output that will influence price. As its output increases, it must lower its price to induce households to purchase the larger quantity. The demand curve for such a firm is shown in Figure 3–2. Notice that when the firm lowers its price in order to sell the increased output, it must reduce the price on *all* units. The change in revenue, for some change in output, therefore, is not reflected only in the unit price but includes the loss on all units sold when the price changes. Algebraically,

$$MR = \frac{\Delta TR}{\Delta q} = p + \left(\frac{\Delta p}{\Delta q}\right)q$$

where $\Delta p/\Delta q$ is the change in price associated with a change in output. Since the demand curve is negatively sloped, $\Delta p/\Delta q$ is negative and marginal revenue is less than price. As shown in Figure 3–2, for every level of output, marginal

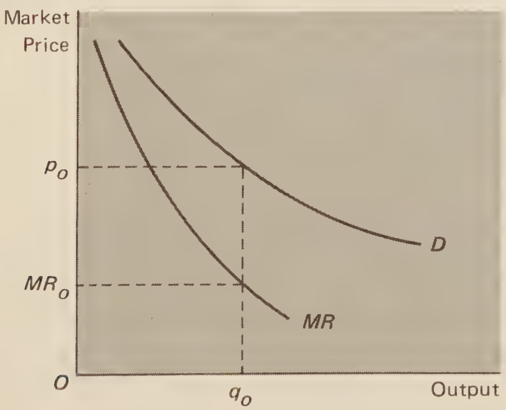


FIGURE 3–2
Relation Between Market Price and Output for a Large Firm (Monopoly Model)

revenue lies below selling price (read from the demand curve). When output is q_o, selling price is p_o and marginal revenue is MR_o.

We noted earlier that a firm maximizing profit will set marginal revenue equal to marginal cost, that is,

$$MR = MC$$

Figure 3–3 shows that a firm large enough to affect market price by its output decisions (in this case a monopoly) will produce q_m (that output for which $MR = MC$) and charge a price p_m, the price at which q_m can be sold. Notice that the Pareto efficient output, q_e, is determined where the demand curve crosses the MC curve, since for q_e selling price p_e will equal marginal cost. Thus, if marginal revenue is less than price, q_m will be less than q_e, the Pareto efficient output, and p_m will be greater than p_e and greater than marginal cost.

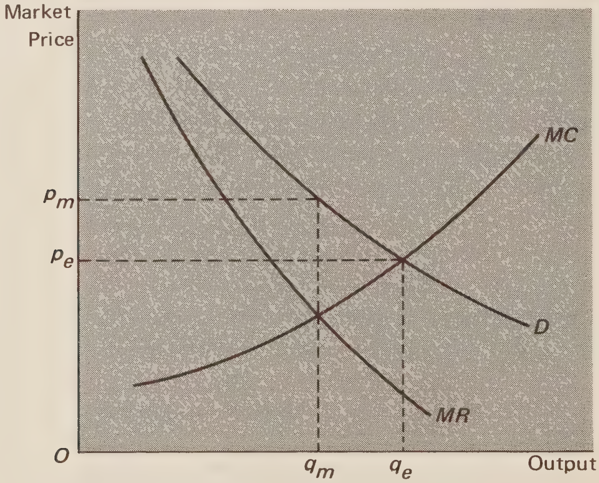


FIGURE 3–3
Monopoly Market Performance

Externalities

Externalities occur when there is a discrepancy between private and social benefits or costs associated with the consumption or production of some good or service.[4] When externalities exist, actions that would maximize private gains or minimize private losses are not generally consistent with social optima. Although

[4] The term *externality* refers to the fact that some good (external economy) or bad (external diseconomy) is generated by the consumer or producer that does not affect him or his decision, that is, the effect is external to his own welfare. Externalities, however, affect the welfare of others.

externalities represent a gain (benefit) or loss (cost), they are not taken into account by market participants and hence are not included in market values.

Consider the case of a rendering plant that produces tallow and emits an objectionable odor. For the owner of the plant the single output is tallow. For society there is a joint product, tallow and odor. Suppose, for simplicity, there is a single variable factor of production, labor. Then the private production function is

$$T = f(L)$$

where T is tallow output and L is labor input. The private marginal product of labor is

$$MP_L^P = \frac{\Delta T}{\Delta L}$$

The social production function is

$$O + T = g(L)$$

where O is odor output.[5]

Now suppose that the odor emitted depends on the level of tallow production, that is,

$$O = h(T)$$

Then the social marginal product of labor is

$$\begin{aligned} MP_L^S &= \frac{\Delta(T+O)}{\Delta L} \\ &= \frac{\Delta T}{\Delta L} + \frac{\Delta O}{\Delta L} \\ &= \frac{\Delta T}{\Delta L} + \left(\frac{\Delta O}{\Delta T} \times \frac{\Delta T}{\Delta L}\right) \\ &= \frac{\Delta T}{\Delta L}\left(1 + \frac{\Delta O}{\Delta T}\right) \\ &= MP_L^P\left(1 + \frac{\Delta O}{\Delta T}\right) \end{aligned}$$

The ratio $\Delta O/\Delta T$ is called the **interaction effect.** The greater the interaction, the greater the discrepancy between marginal social product and marginal private product of labor.

We know that the firm will hire workers until the value of the marginal product of labor is equal to the wage. If p_T is the market price of tallow, this implies

$$MVP_L^P = p_T \times MP_L^P = w$$

[5] The unit in which O is measured is immaterial, provided a link is established between O and social welfare.

where MVP_L^P is the private marginal value product of labor. For a social optimum, the value of the marginal social product of labor must equal the wage. If p_O is the social value of a unit of odor, then the value of the marginal social product, MVP_L^S, is

$$MVP_L^S = p_T \frac{\Delta T}{\Delta L} + p_O\left(\frac{\Delta O}{\Delta T} \times \frac{\Delta T}{\Delta L}\right)$$

$$= p_T \frac{\Delta T}{\Delta L}\left[1 + \left(\frac{p_O}{p_T} \times \frac{\Delta O}{\Delta T}\right)\right]$$

$$= MVP_L^P\left[1 + \left(\frac{p_O}{p_T} \times \frac{\Delta O}{\Delta T}\right)\right]$$

and the social optimum occurs where

$$MVP_L^S = MVP_L^P\left[1 + \left(\frac{p_O}{p_T} \times \frac{\Delta O}{\Delta T}\right)\right] = w$$

Since odor represents a loss in utility, p_O must be negative, so that

$$MVP_L^S < MVP_L^P$$

Figure 3–4 shows the social and private demand for labor in tallow production and the associated private and social optimal levels of employment for a fixed wage. When the wage is w, L_p units will be employed. However, L_s is the socially optimum level of employment in tallow production. Thus private decisions result in overemployment in the tallow industry and an oversupply of tallow.

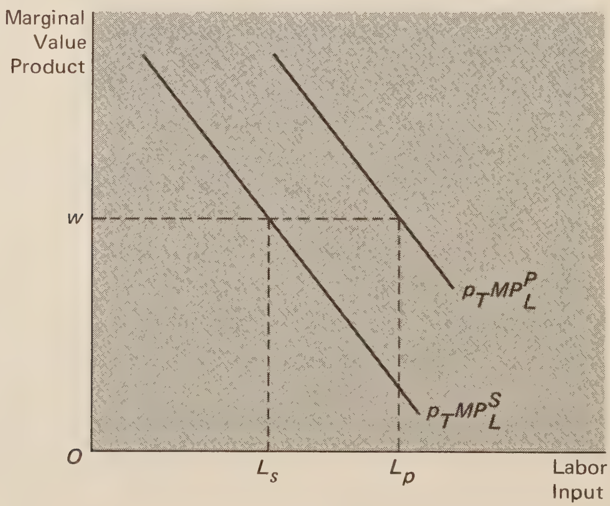


FIGURE 3–4
Private and Social Employment Optima in the Presence of an External Diseconomy

Suppose the production process generates a good rather than a bad. Professor James Meade cites the example of an apple farmer who generates an externality to a neighboring beekeeper in the form of nectar.[6] Bees produce more honey, the more nectar they obtain from apple blossoms. In this case, the marginal social value product of labor in apple production, MVP^S_{LA}, is

$$MVP^S_{LA} = MVP^P_{LA}\left[1 + \left(\frac{p_H}{p_A} \times \frac{\Delta H}{\Delta A}\right)\right]$$

where p_A is the price of apples, p_H is the price of honey, MVP^P_{LA} is the marginal private value product of labor in apple production, and $\Delta H/\Delta A$ is the interaction effect on honey production of apple production.

Since p_H and p_A are greater than zero, the marginal social product of labor in apple production exceeds the marginal private product. As shown in Figure 3–5,

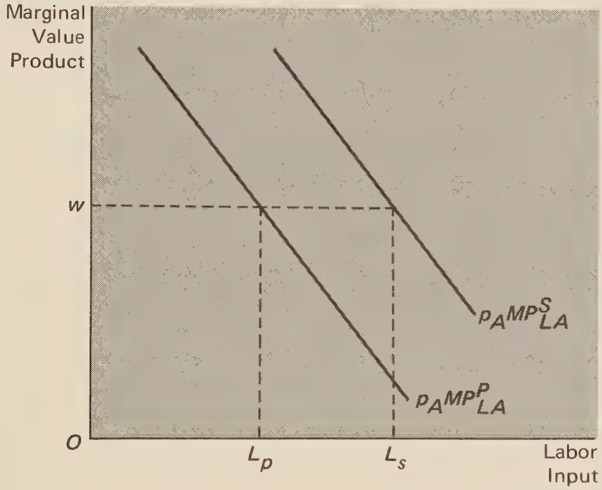


FIGURE 3–5
Private and Social Employment Optima in the Presence of an External Economy

private profit maximization results in less labor employed in apple production, L_p, than is socially optimal, L_s, and a consequent underproduction of apples.

Another type of externality occurs when there is a negative interaction effect on one good when another is produced. One frequently cited example is the factory smokestack that belches soot onto the clean linen of the neighboring laundry.

In all these cases, there is a discrepancy between social and private costs and benefits associated with a particular activity. Since market values reflect private

[6] *Economic Journal, LXII* (March 1952), pp. 54–67. This example is also cited in Francis M. Bator, "The Anatomy of Market Failure," *Quarterly Journal of Economics, LXXII* (August 1958), pp. 351–379.

costs and benefits, they understate (in the case of activities generating external economies) and overstate (in the case of activities generating external diseconomies) true social values. If a price system could be devised that would reflect social values, it would produce optimal social decisions. An attempt to devise such a price system is at the heart of tax-subsidy schemes that would purportedly remedy the effect of externalities on market decisions.

Suppose we were to place a per unit tax t on tallow production such that

$$t = -p_O \times \frac{\Delta O}{\Delta T}$$

where p_O is the social value of the odor produced by the rendering plant, O is the odor output, and T is the tallow output. If such a tax were imposed, the private marginal value product of labor in tallow production would be

$$MVP_L^P = (p_T - t)\frac{\Delta T}{\Delta L}$$

$$= p_T \frac{\Delta T}{\Delta L}\left[1 + \left(\frac{p_O}{p_T} \times \frac{\Delta O}{\Delta T}\right)\right] = MVP_L^S$$

That is, the private marginal value product would exactly equal the social marginal value product and hence

$$MVP_L^P = MVP_L^S = w$$

Optimal private decisions would coincide with optimal social decisions.

Equivalently, subsidies of an amount

$$S = p_Y \times \frac{\Delta Y}{\Delta X}$$

where p_Y is the social value of Y, and $\Delta Y/\Delta X$ is the interaction of Y with X, could be provided in the cases of external economies, where Y, a good, is produced external to the production of X.

Public Goods

The most extreme case of an externality is a pure **public good.** Samuelson defines a pure public good as one for which total social consumption is the same as each person's consumption. If Y is total consumption of a public good, then

$$Y = Y_1 = Y_2 = \cdots = Y_n$$

where Y_i is consumption of the good by the ith individual.[7] National defense and programs to eliminate air pollution are examples of public goods and services

[7] Paul A. Samuelson, "Diagrammatic Exposition of a Theory of Public Expenditures," *Review of Economics and Statistics, XXXVII* (November 1955), pp. 350–356.

available to everyone in equal amounts. Cleaner air benefits everyone, and one person cannot increase his consumption of clean air at another's expense.

On the other hand, total social consumption of a private good, X, is the sum of each person's consumption of it, that is,

$$X = Y_1 + X_2 + \cdots + X_n$$

where X_i is consumption by the ith individual. If one person consumes a private good such as food, clothing, or an automobile, he prevents another from consuming it.

Mixed public goods partake of some aspects of public and private goods. Although many people benefit from the provision of these goods, the amount one consumes may affect another's ability to consume the good. Education, public transportation, spectator sports activities are mixed public goods. Two people can enjoy a lecture, but its effectiveness is reduced if the room becomes very crowded. Thus the entry of one more person may reduce the benefits to others in the room. For simplicity we will assume that goods are either purely public or purely private. The application to mixed public goods follows directly from this analysis.

When a private good is produced, it is consumed by a single individual. Thus that individual's private gain from consuming that good, reflected in his marginal rate of substitution of that good for other goods, is equal to the social gain. Assuming the social cost is reflected in the marginal rate of transformation (there are no production externalities), it follows that a social optimum occurs where

$$MRS = MRT$$

and where

$$MRS_1 = MRS_2 = \cdots = MRS_n$$

MRS_i being the marginal rate of substitution of the public good for other goods for the ith consumer.

For public goods, which everyone consumes equally (your war is my war, and so on), the social gain from the production of one or more unit is the *sum* of the marginal rates of substitution for all individuals in the economy. Thus it follows that allocation of resources to public goods is socially optimal where

$$MRS_1 + MRS_2 + \cdots + MRS_n = MRT$$

For optimal allocation of resources to public goods the sum of the individual marginal rates of substitution must be equated to their marginal rate of transformation. Consequently, a private market system in which individuals adjust their purchases and producers adjust production so that

$$MRS_{YX} = \frac{p_X}{p_Y} = MRT_{YX}$$

will necessarily underproduce Y. That is, private decisions will result in less public Y and more private X than is socially optimal. In terms of the externality discussion, social value exceeds private value to the greatest possible extent. That is why public goods can be viewed as the extreme externality.

A price system alone is not adequate to ensure optimal provision of public goods. Neither is the sort of generalized subsidy scheme applicable in the case of production externalities.[8] Notice that each individual at the margin has the same subjective rate of trade-off of a private good for other goods. Thus, charging fixed prices (or prices plus fixed subsidy) for the good will be sufficient to produce this result. However, individuals can and generally do consume public goods where, at the margin, the subjective rate of exchange with private goods varies among consumers. Since all individuals must consume equal amounts of the public good, Y, then on the average the marginal rate of substitution of public for private goods (the amount of public goods required to "compensate" the consumer for giving up private goods) will be lower for those consuming more private goods, the rich. If consumer I has more private goods than consumer II, that is,

$$X_{\mathrm{I}} > X_{\mathrm{II}}$$

it is likely that

$$MRS^{\mathrm{I}}_{YX} < MRS^{\mathrm{II}}_{YX}$$

Consumers will be satisfied with the amount of public goods provided as long as

$$MRS_{YX} = \frac{p_X}{p_Y}$$

Since individuals have different MRS_{YX} for any level of Y, then unless individuals are charged different prices for public goods, efficient allocation of resources to public goods will not occur. If MRS_{YX} declines, on the average, with the level of income, taxes should be higher at higher income levels.[9] However, for true efficiency each household should be charged a different price for public goods, since MRS_{YX} will vary among households depending upon individual tastes as well as income levels. To establish such a price system would be a practical impossibility, since it requires a knowledge of the subjective preferences of each consumer in the economy. This was not the case for private goods, for which fixed prices ensured that the efficiency condition

$$MRS^{\mathrm{I}}_{XY} = \frac{p_Y}{p_X} = MRS^{\mathrm{II}}_{XY}$$

[8] The external effect in the case of public goods comes from the consumption side. The examples in the previous section were of production externalities.

[9] Here is an argument for income taxation based on the criterion of efficiency alone. Arguments for income taxation are generally concerned with distributive questions. While most people can agree on the desirability of efficiency, there is likely to be considerable disagreement on the optimal income distribution, and consequently arguments based on efficiency are often more powerful than those based on distributional equity.

would be satisfied. Consumers will automatically adjust their purchases to satisfy the efficiency condition.

If a price system were used to assure efficient allocation of resources to public goods, consumers would not reveal their preferences or, equivalently, the amount they would be willing to pay for public goods, since they could enjoy the full amount of Y regardless of how much they paid for it. To the extent that some public goods are not *pure* public goods (for example, highways, schools), this statement must be qualified. In all cases, however, the social benefit exceeds the benefit to the individual, and hence the individual cannot be forced to pay the full amount of what he enjoys.

To summarize, given the discrepancy between social benefit and individual private benefit associated with public goods, a market system will underproduce public goods. Furthermore, any attempt to finance public goods on the basis of prices charged to users would fail. Consequently, the public sector finances such activities from tax revenues raised on the basis of other criteria.[10]

Income Distribution

A Pareto optimum is attained when no one can be made better off without making someone else worse off. Since there is a different Pareto optimum associated with every conceivable distribution of welfare, and assuming a direct relation between welfare and income, the Pareto criterion clearly skirts the issue of the optimal income distribution.[11]

Figure 3–6 shows a utility trade-off frontier between two individuals or, alternatively, between one individual and everyone else taken together. Points on the frontier, such as A, represent a Pareto optimum, since U_{I}, the utility of individual I, can only be increased at the expense of the utility of individual II, U_{II}. On the other hand, points inside the frontier, such as B, are inefficient, since U_{I} can be increased by moving to A without reducing U_{II}, that is, without making individual II worse off. Alternatively, a move from B to C would make II better off without making I worse off.

The Pareto optimum can be viewed as the best allocation of resources *for any given income distribution*. The question of the "best" income distribution, however, requires the sort of interpersonal value judgment associated with robbing Peter to pay Paul. Consequently, many economists have considered such judgments beyond their scope. Others base criteria for optimal income distribution on theoretical arguments.

For instance, if we assume that for higher income levels the gain in utility

[10] Criteria most commonly put forward for taxation are neutrality with respect to the price system, distributional equity, ability to pay, and the benefit principle. For a discussion see Richard A. Musgrave, *The Theory of Public Finance* (New York: McGraw-Hill, 1959), chapters 4 and 5.

[11] Economic welfare increases with the consumption of goods as well as of leisure. If low incomes are associated with a high consumption of leisure, then there is not necessarily a positive relation between income and welfare. If a measure of income is to reflect welfare, it should include leisure. Also, there are welfare effects associated with different types of work that should be taken into account.

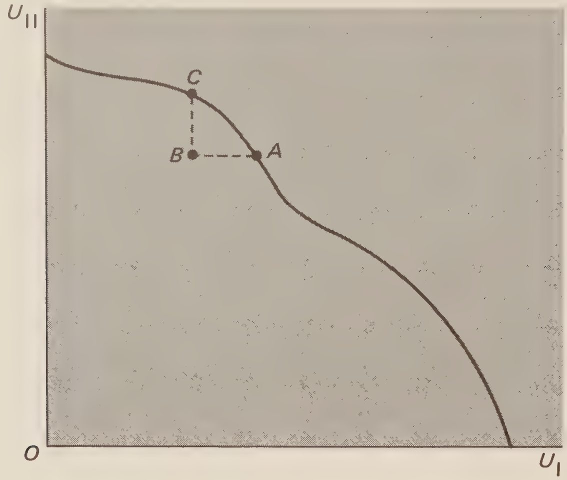


FIGURE 3–6
A Utility Trade-Off Frontier

associated with a $1 increment in income is lower than at low levels of income, then redistribution of income toward greater equality is viewed as desirable on the basis of the principle of maximizing social welfare. Such a conclusion still requires interpersonal comparisons, however. Although for each individual the marginal utility of income may fall, as in Figure 3–7,[12] it is not necessarily true that the marginal utility of income for some high-income individual (or for that matter, all high-income persons taken together) is any smaller than that for low-income individuals. The latter involves comparing the psychic satisfaction experienced by different individuals.

Income Distribution in a Market System

A competitive market model, in the absence of externalities, will produce a Pareto optimal allocation of resources, but that allocation may produce a distribution of income that is viewed (at least by some observers) as undesirable. In a market model the return to factors of production will equal their marginal value product as perceived by their employers.

Firms will employ productive factors until their marginal value product equals the factor price, that is, until

[12] It is not obvious that the marginal utility of income should decline at higher income levels. We generally assume that the marginal utility of consuming a *single good* falls as more is consumed, because of a "saturation effect." But it doesn't necessarily follow that given the large variety of goods and services available, marginal utility will fall as more of all of these is consumed. Marginal utility of income could conceivably rise as the life style improves, at least up to a certain point. A welfare family, discouraged and pessimistic, may incur less psychic satisfaction from a $100 windfall than the same family in more affluent days, when it is upwardly mobile, with aspirations of achieving middle-class status.

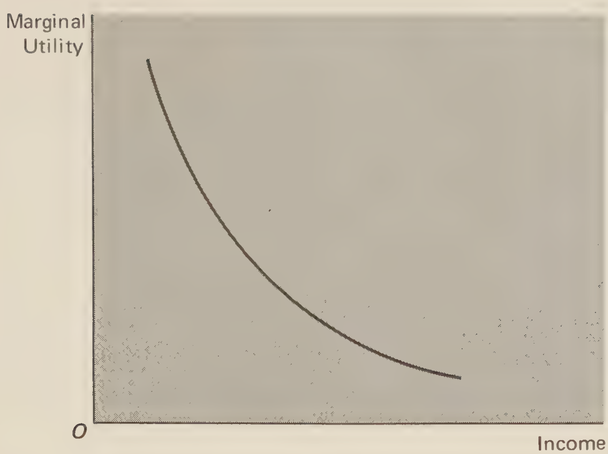


FIGURE 3-7
Diminishing Marginal Utility of Income for an Individual

$$MVP_f = p_f$$

For an individual household income is determined by the amount of productive factors owned and supplied to factor markets as well as by the productivity of those factors. Consider a household that owns land and capital in addition to participating in the labor force. Income for that family can be expressed as

$$Y_i = w_iL_i + r_iK_i + a_iA_i$$

where w_i is the wage earned by the ith household, r_i is the return to its capital, a_i is the return to its land, L_i is the amount of labor provided by the household, K_i is the capital owned by the household, and A_i is land used by the household for productive purposes. Alternatively,

$$Y_i = (MVP_{Li})L_i + (MVP_{Ki})K_i + (MVP_{Ai})A_i$$

where MVP_{ji} is the marginal value product of the jth factor for the ith household.

Differences in income among households will depend upon the amount of productive factors they own and supply to factor markets as well as upon the productivity of the factors supplied. Both of these conditions are largely influenced by social, legal, and economic institutions. Inheritance laws and customs as well as tax policy affect property ownership. Yet even if property were equally distributed, inequality would remain because of differences in factor productivity, particularly with respect to labor. In addition, there may be individual differences in the amount of owned factors that are actually supplied to factor markets.

Some economists argue that differences in labor productivity reflect differences in effort and that consequently income distribution in a market system is a "just" reward for such "effort." However, productivity differences are also

attributable to social and cultural factors such as inequality of educational opportunity, social class and ethnic traditions, and differences in capacity to acquire "human capital," which are more or less beyond the control of the individual. Furthermore, distinctions as to productivity based on sex, age, and race that are perceived by employers may not reflect true differences. As an economy moves from a goods orientation to emphasis on services, the notion of productivity becomes increasingly open to subjective interpretation on the part of the employer.

Clearly, the issue of the optimal income distribution is open to debate. Furthermore, the market model suggests that distribution in a market system depends on factor ownership and productivity, which is not necessarily the most equitable arrangement. There is, however, an efficient allocation of resources associated with every distribution of income, and consequently, associated with the optimal income distribution there is an efficient allocation of resources. If the optimal income distribution is attained, then the overall social optimum can only be reached when Pareto efficiency is achieved. In evaluating economic institutions in terms of their ability to produce efficiency once the optimal income distribution is achieved, it is important also to examine their flexibility in facilitating socially desirable redistributions of income.

To those who have relied on laissez faire to provide efficient resource allocation in a market system, it is uncomfortable to have to accept an explicit decision concerning the "for whom" aspect of the allocative question. However, the issue must be faced, since it is unlikely that the distributive consequences of laissez faire will be found satisfactory, at this time, to any but its most fortunate beneficiaries.

3-3 THE THEORY OF SECOND BEST

Given the possibilities for the failure of a market system to produce an optimal allocation of resources, we may ask what remains of the market model as a guide for public policy. The theoretical justification for a market system is that given the assumptions of the idealized market model, the Paretian marginal conditions are satisfied. What happens to the model if one or more of the assumptions are violated in an actual market system or, alternatively, if one or more of the Paretian conditions are not satisfied?

At first glance, the theory of second best strikes a death blow to the use of the market model as a guide to public policy.[13] The theory of second best states that if one of the Paretian marginal conditions is violated in an economy, the other marginal conditions, although still attainable, are in general no longer desirable. Further, it is not true that a situation in which more but not all of the

[13] See R. G. Lipsey and Kelvin Lancaster, "The General Theory of Second Best," *Review of Economic Studies, XXIV* (1956–57), pp. 11–32.

Paretian conditions are fulfilled is likely to be superior to a situation in which fewer are fulfilled.

This makes sense, of course. Each marginal condition was derived under the assumption that the others were satisfied. Suppose, for instance, that the steel industry is a monopoly but that in all other industries firms are so small that their decisions have a negligible effect on market prices. Since steel will be sold at a price greater than its marginal cost, that is,

$$p_S > MR_S = MC_S$$

the marginal rate of substitution of other goods for steel will exceed the marginal rate of transformation, that is,

$$MRS_{GS} = \frac{p_S}{p_G} > MRT_{GS}$$

This implies that steel is underproduced relative to other goods or, equivalently, that other goods are overproduced relative to steel. If the steel industry remains a monopoly (because of economies of scale or institutional factors), it is no longer desirable for firms in other industries to produce where price equals marginal cost. Instead, they too should sell at a price that exceeds marginal cost.[14] Thus, the "second best" resource allocation requires that all firms violate the principle of marginal cost pricing.

A similar analysis can be made in the presence of externalities. If production in one firm or industry deviates from the social optimum in the presence of externalities, then presumably the attainment of private optima that are consistent with social optima in other firms and industries (that is, there are no externalities) is no longer desirable.

Does the theory of second best imply that microeconomic theory cannot provide a guide to policies for optimal resource allocation? No—provided microeconomic analysis is applied correctly. We can use microeconomic analysis to determine the new Paretian conditions, that is, the conditions that will produce Pareto efficiency given the constraint imposed by the situation. For instance, given the constraint that the steel industry is a monopoly, we apply microeconomic theory to find the equilibrium conditions in which no change can be made that makes someone better off without making someone else worse off.

The theory of second best merely warns against the blind application of the Paretian marginal conditions to problems of resource allocation. The marginal conditions are not ends in themselves but means to an end—efficient resource allocation. And they are only means to that end in the case in which they can all be satisfied. If one or more are violated, new conditions must be derived.

[14] This is not necessarily the point at which marginal revenue equals marginal cost, either. The relation between marginal revenue and price varies among industries depending on the shape of the respective demand curves. Efficiency occurs when the ratio of selling prices for two industries equals the ratio of their marginal costs.

It is valid, however, in view of the theory of second best, to question policies that attempt to impose the Paretian marginal conditions on a piecemeal basis, that is, to individual industries and firms in isolation. Should we, for instance, regulate the aluminum industry and force it to price at marginal cost when the steel industry remains an unregulated monopoly? The answer may depend largely on the degree of interrelation between the goods involved.[15] The more closely the goods are interrelated from the viewpoint of the consumer (as are steel and aluminum) or of the producer (as are provision of electric power to domestic versus commercial users), the less desirable are piecemeal policy measures applied to one of the goods. On the other hand, the more unrelated the products, the more negligible the impact of a violated Paretian condition in one industry on the other. Policies to promote competition in the service station industry, for instance, may be applied without regard to conditions in the industry producing toothpaste.

3–4 TAXATION AND ECONOMIC EFFICIENCY

The provision of public services requires the levying of taxes. One of the most generally accepted principles of taxation in a market economy is that taxes should be neutral with respect to the price system. Economists generally favor income or sales taxes over selective excise taxes on individual goods.

Excise taxes are viewed as undesirable because of the presumed efficiency of a system of market prices. In the market model, if producers and consumers sell and buy at fixed market prices, this ensures that the marginal rate of product substitution between any two goods will equal the marginal rate of transformation, that is,

$$MRS_{YX} = \frac{p_X}{p_Y} = MRT_{YX}$$

If a tax, t, is imposed on X (and assuming that the tax is paid by the consumer), the effective price to the consumer becomes $p_X(1 + t)$. Consumers will purchase both goods until the marginal rate of product substitution is equal to the ratio of the effective prices, that is,

$$MRS_{YX} = \frac{p_X(1 + t)}{p_Y}$$

Producers will continue to produce until

$$\frac{p_X}{p_Y} = MRT_{YX}$$

[15] For a theoretical demonstration of this point see James M. Henderson and Richard E. Quandt, *Microeconomic Theory*, 2d ed. (New York: McGraw-Hill, 1971), pp. 286–288.

so that

$$MRS_{YX} = \frac{p_X(1+t)}{p_Y} > \frac{p_X}{p_Y} = MRT_{YX}$$

The tax, in effect, drives a wedge between buying and selling prices so that *MRS* will exceed *MRT*. Consequently, the taxed good is underproduced.

Selective excises are not the only type of non-neutral tax. Income taxes are essentially subsidies to leisure (since they reduce the opportunity cost of leisure in terms of income) that will interfere with the optimal supply of labor. In the absence of such taxes households will set the marginal rate of substitution of income for leisure equal to the wage. Firms will hire labor until the marginal dollar product equals the wage, that is,

$$MRS_{Yl} = w = MP_L(\$)$$

This satisfies the efficiency condition for optimal labor supply.

If an income tax, t, is imposed, the effective wage (from the viewpoint of the household) becomes $w(1-t)$. Then households will set

$$MRS_{Yl} = w(1-t)$$

while firms continue to set

$$w = MP_L(\$)$$

Thus

$$MRS_{Yl} = w(1-t) < w = MP_L(\$)$$

that is, the marginal rate of substitution of income for leisure is no longer equal to the opportunity cost of leisure in terms of income.

A similar analysis can be applied to sales taxes, which penalize consumption. Such taxes affect the rate of saving (the preference for present over future consumption) and hence the optimal supply of capital. It is difficult, if not impossible, to devise a totally neutral tax.

The case for neutrality in taxation rests on the assumption that the existing price system is producing efficient resource allocation in the first place. If, in fact, there are externalities or monopolies, selective excises could be justified. In the discussion of externalities it was noted that a tax (or subsidy) could be devised that would induce producers to attain the social rather than the private efficient allocation of resources. A similar case could be made for subsidizing monopolies that underproduce relative to industries that follow marginal cost pricing.[16]

In fact, proponents of a market system as the best way to achieve efficient

---

[16] Subsidizing monopolies will undoubtedly have undesirable distributive consequences, which will have to be rectified by other forms of taxation.

resource allocation have traditionally advocated the use of selective excises to remedy the inefficiencies produced by externalities and monopoly.[17] Thus a tax on gasoline may be justified on the grounds that gasoline engines produce undesirable air pollution. Gasoline taxes can also be justified as "user taxes" for public highways and are an attempt to extract a payment from private individuals for a public good.

Taxation also has distributive consequences even if the intent is not purely redistributional. A tax on cigarettes, for instance, is clearly regressive. Since a market system does not necessarily produce an optimal income distribution, the redistributive aspects of a tax system may or may not be desirable. But in the absence of any theoretical criteria for optimal distribution of income, the evaluation of the distributive consequences of taxation is largely a matter of personal judgment.

QUESTIONS FOR STUDY AND REVIEW

1. What is the theoretical justification for a market system composed of small consuming and producing units? What are some of the weaknesses of this theoretical rationale?
2. Economists generally favor income or sales taxes over selective excise taxes on individual goods. Present and evaluate critically the theoretical justification for such a position.
3. A factory smokestack produces soot, which has forced a neighboring laundry to install an expensive ventilating system. The laundry takes the factory to court, demanding that the factory owners bear the cost of installing the new equipment. The lawyers for the factory owners contend that no damages should be awarded, since the laundry firm is free to relocate to a "cleaner" part of the city. If you were the judge, how would you rule?
4. Discuss the welfare implications of the following schemes for financing education:
   a) Both private and public education should be self-financing, and families should pay the cost of their children's education.
   b) All citizens are taxed in relation to their income and wealth. Children can receive free education in public schools up to the compulsory minimum, the cost being financed out of taxes. Private schools are self-financing, and families must pay the full cost of their children's education in such schools.
   c) Citizens whose children attend public schools are taxed according to their income and wealth. Any child may receive free education in public schools up to the compulsory minimum, the cost being financed out of taxes. Private schools are self-financing, and families must pay the full cost of their children's education in such schools.

---

[17] See, for instance, Alfred Marshall, *Principles of Economics* (London: Macmillan and Co., 1959), book V, chapter 12; and A. C. Pigou, *The Economics of Welfare*, 3d ed. (London: Macmillan and Co., Ltd., 1929).

5. What are the implications of the theory of second best for microeconomic theory? For microeconomic policy in a market system?
6. A petroleum refinery emits thirty tons of sulfur oxide per $100,000 of output. Sulfur oxide is estimated to have a "nuisance cost" to society of $20 per ton. The production function for the firm is shown below:

Labor Input (Man-Weeks)	Petroleum Output ($1,000 of Value)
1,000	500
2,000	900
3,000	1,200
4,000	1,400
5,000	1,500

   a) If the going wage is $200 per week, how many workers will be employed?
   b) What is the marginal private product?
   c) What is the marginal social product?
   d) Can you devise a tax on petroleum that will induce the firm to produce at the social optimum? What will be the effect of this tax on employment for the firm? On output?
   e) How would your analysis change if the "nuisance cost" of sulfur oxide increased to $100?

ADDITIONAL READING

Bator, Francis M. "The Anatomy of Market Failure," *Quarterly Journal of Economics, LXXII* (August 1958), pp. 351–379.

Baumol, W. J. *Welfare Economics and the Theory of the State,* 2d ed. London: G. Bell, 1965.

Friedman, Milton. *Capitalism and Freedom.* Chicago: University of Chicago Press, 1962.

———. "The 'Welfare' Effects of Taxes," in Milton Friedman, *Essays in Positive Economics.* Chicago: University of Chicago Press, 1953, pp. 110–113.

Mishan, E. J. "Reflections on Recent Developments in the Concept of External Effects," *Canadian Journal of Economics and Political Science, XXXI* (February 1965), pp. 1–34.

Phelps-Brown, E. H. *The Framework of the Pricing System.* London: Chapman & Hall, 1936.

4 Alternative Economic Systems

One of the great controversies of economics concerns the question of whether or not resources can be allocated efficiently in the absence of markets. In this chapter we develop a theoretical model of non-market planning that produces Pareto efficient resource allocation. Like the market model, however, this model of competitive socialism is a theoretical construct, with assumptions that are never found to be completely true in the real world. As with the market model, the evaluation depends largely on the implications for a real-world economy in which some, but not all, of the underlying preconditions set forth in the model are met.

4–1 RESOURCE ALLOCATION OUTSIDE THE MARKET

We have examined a theoretical market model as a system for providing efficient allocation of resources. Although in theory such a model produces efficiency, except where externalities are present, it is clear that even in free enterprise capitalism all goods and services are not traded in markets such as the model describes.

Modern technology in many industries is characterized by substantial economies of scale, that is, unit costs are lowest when production is undertaken on a large-scale basis. Since larger firms can profitably sell their output at lower prices than small firms, the latter are driven out of the market until it is dominated by a few corporate giants. It was noted that once a firm is large enough to influence market price, its pricing and output decisions are no longer consistent with efficient resource allocation, even in a market system. Also, such giant enterprises generally produce a multiple product line, employ many factors of production, and often monpolize sources of factor supply. Consequently, there do not always exist reliable indicators of the opportunity cost of acquiring resources or the marginal rates of transformation between various products. Within a large enterprise allocative decisions—what is to be produced and how it is to be produced—must be made in the absence of markets.

The public sector in the United States economy controls about 25 per cent of the nation's resources. It is in the nature of public goods that market prices (as determined by private consumption demand) are not accurate indicators of the social demand for public goods. Thus the question of *what* is to be produced cannot be determined in the market. Some public goods are acquired directly from private producers—military aircraft, building projects contracted to private firms—and hence the question of *how* these public goods are to be produced is determined in the private sector (which may or may not rely on market forces, depending on the size of the firm). However, other public goods are produced in circumstances in which the government controls factor supply and determines how factors are utilized. National defense, the provision of administrative services by the federal bureaucracy, the use of public lands all require nonmarket decisions concerning *how* goods and services are to be produced.

There are other reasons for abandoning the market as a mechanism for allocating resources. Socialists claim that since private ownership of capital is inherent in a capitalist system, equitable redistribution of income can never be accomplished. This is because political power belongs to the capitalist elite, which will never allow more than token redistribution. Furthermore, a class society controls the way in which individuals can acquire human capital and productive capacity and hence can perpetuate inequalities for generations.

Most socialist countries[1] (Yugoslavia being a notable exception) have abandoned the use of markets as a means of allocating resources. The reasons for this seem to be based on the historical relation between the market system and capitalist ownership of means of production. Then, too, if society is to take advantage of technological economies of scale, it must organize into large enterprises. But large-scale industrial organization is not consistent with efficient resource allocation in a market system. Finally, it is generally accepted in a socialist state that social goals may supercede private wants and that therefore the state must be able to indicate priorities with respect to *what* is to be produced (to the extent that these are inconsistent with private goals). In contemporary socialist countries the emphasis has been on growth rather than on static efficiency. Consequently, these societies have been willing to exchange consumer sovereignty in the short run for faster rates of growth that will benefit future generations.

Given these conditions, it is clear that all societies, whether socialist or capitalist, must face the question of efficient resource allocation in a nonmarket context. And in the perennial ideological debate between capitalism and socialism it is important to have a basis for comparing market institutions with nonmarket institutions in terms of the efficiency criteria.

[1] By *socialist* I mean countries that have totally abandoned private ownership of capital and that have an explicitly anticapitalist ideology. Some countries, such as Sweden, are often labeled socialist even though they retain private ownership of capital.

4-2 CENTRAL PLANNING: THE TARGET APPROACH

Having derived the criteria for achieving economic efficiency—the Paretian marginal conditions or their extensions (in the case of second best solutions)—we can theoretically determine the optimal resource allocation for any economy. If we knew (or estimated) every individual's utility function, the production functions for each firm, and the conditions of factor availability or supply, and if we put all this information into a giant computer and programmed it to find the resource allocation that satisfies the marginal conditions, a solution could be obtained that would be independent of the institutional arrangements of the society.[2]

Now, of course, obtaining such a solution and constantly revising it on the basis of changing tastes and technology would be an enormous job. In addition to the task of acquiring the information to achieve a solution, a great bureaucracy would be required to ensure that the proper allocation was taking place. Each firm would be allotted specified amounts of factor services with a stipulated output target, and workers would be assigned occupations and positions within firms with a stipulated number of working hours required. Consumer goods would be distributed on the basis of ration tickets allotted to each family. It is difficult to imagine how all of this could be achieved in a static economy, much less one in which tastes and technology are constantly changing!

In addition to the practical difficulties involved in centrally planning the allocation of resources, there is a basic theoretical problem with noncapitalist forms of organization. If capital is not privately owned, there can be no way to determine the value of capital goods. In a capitalist economy capital value is determined in part by consumer preferences for present over future consumption. If capital has no value, there is no way to determine the optimal utilization of productive factors.

After the establishment of a socialist state in the Soviet Union in the 1920s, several Western economists, von Mises, Hayek, and Lionel Robbins, pointed out both the logical and practical impossibilities of centrally planned resource allocation. A market system, it was argued, is the only method for performing the necessary calculations and corresponding allocations efficiently. The market mechanism provides all the information for making efficient allocative decisions quickly, without the requirement for explicit calculation. Furthermore, as has been noted, once market prices are established, all the individual market participants will automatically make efficient choices without the need for allotting specific inputs and target outputs or for rationing.

The socialist response to this was the development of a model of decentralized socialism that would in theory produce efficient resource allocation without requiring the information and explicit bureaucratic guidance inherent in the

---

[2] Except, of course, to the extent that utility functions and technology are themselves in part institutionally determined.

target output approach. Oskar Lange, who first developed the complete model in his *On the Economic Theory of Socialism,* published in 1938, acknowledged his debt to von Mises, the most vocal critic of nonmarket forms of organization:

> . . . The merit of having caused the socialists to approach the problem systematically belongs entirely to Professor Mises. Both as an expression of recognition for the great service rendered by him and as a memento of the prime importance of sound economic accounting, a statue of Professor Mises ought to occupy an honorable place in the great hall of the Ministry of Socialization or of the Central Planning Board of the socialist state. I am afraid, however, that Professor Mises would scarcely enjoy what seems the only adequate way to repay the debt of recognition incurred by the socialists, and it is difficult to blame him for not doing so.[3]

Von Mises pinpointed the impracticality of target planning for resource allocation. Not only are the costs of information and implementation prohibitive but social ownership of the means of production posed theoretical questions concerning criteria for optimal resource utilization. By formulating the problem in this way, von Mises forced the development of a new model for a socialist economic system. As we will see, that model is applicable to nonmarket resource allocation in a capitalist system as well, since it does not depend upon who owns or controls the means of production.

4–3 PRICES AND ECONOMIC EFFICIENCY

It was mentioned in chapter 2 that with every distribution of resources satisfying the Paretian conditions there is associated a set of prices that, if the resources and goods were traded at those prices, would produce exactly that allocation. This property of the Paretian solution, known as **duality,** is extremely important for the theory of resource allocation. It implies that there exists some price system that would cause all participants in the economy to make efficient allocative decisions. Consequently, if such a system can be established, there would be no need for central enforcement.

A market system is one way to obtain an efficient system of prices. It was noted that (ignoring externalities) the prices established in a market model will produce efficient allocation of resources. To the extent that an actual market system deviates from the market model, as in the case of the natural development of monopolies in the presence of technological economies of scale, the price system established will not produce efficient resource allocation.

The set of prices that will produce efficiency is obtained automatically from the solution to the Paretian problem. That is, if we feed into the computer the

---

[3] Oskar Lange, *On the Economic Theory of Socialism* (Minneapolis: University of Minnesota Press, 1938), pp. 57–58.

utility functions of all households, the production functions of firms, and factor supply equations, and obtain a solution satisfying the Paretian conditions, we will not only know what to produce and how to produce it but also obtain as a by-product the prices that would produce this solution.

The principle of duality demonstrates that the traditional association of prices only with capitalism or with a market system is incorrect. Since a price system is implicit in any structure of resource allocation, any economic system that allocates resources can use a set of prices to achieve its goal. Hayek argues that "we must look at the price system as . . . a mechanism for communicating information if we want to understand its real function." He points out that by raising the price of a single resource we can cause tens of thousands of people to use it more sparingly "without an order being issued." [4]

The idea that a price system could be used for central planning in a nonmarket economy was itself a great step forward. By eliminating the necessity for issuing input quotas and output targets to individual firms and ration tickets to households, the price mechanism decentralizes decision making. Furthermore, it presumably should improve individual morale and production incentives, since it gives the appearance of considerably more individual freedom and autonomy than the target input and output approach. Workers can supply all the labor they wish to any enterprise at a fixed wage rather than being told how long and where they must work. Enterprises can sell all they want to at the established prices, and provided that there is an incentive to make profits (by tying managers' incomes or status to profits), they will select the lowest cost production techniques. Although established prices theoretically produce the same results as output targets (and hence are equally binding on decision makers), they give the semblance of providing more freedom to the participants in the enterprise.

Although decentralized planning in a nonmarket economy is clearly preferable to the target planning approach, the basic problem of finding the optimal price system has not been solved. We still need all of the information concerning tastes and technology to feed into the computer that was required in the target planning approach to obtain an efficient price solution. In a market system efficient prices are obtained automatically. It was the contribution of Oskar Lange to develop a model of competitive socialism, which would produce efficient resource allocation using a price system that would not require the information necessary to the target planning approach.

4–4 THE LANGE MODEL OF COMPETITIVE SOCIALISM

Lange's model incorporates the assumption that efficient resource allocation and the corresponding price system in a socialist state have the same underlying determinants as in a market economy.

---

[4] Friedrich A. Hayek, "The Use of Knowledge in Society," *American Economic Review*, XXXV (September 1945), pp. 526–527.

In a market model efficiency is assured in the following ways:

1. *Efficiency in consumption.* Consumers view market prices as fixed and maximize their utility subject to the limitations of their incomes. This ensures that all consumers have the same marginal rates of substitution (subjective preference) between products.
2. *Efficiency in production.* Firms view market prices as fixed and maximize profits. This ensures that price will equal marginal cost, so the marginal rates of transformation are the same between all products.
3. *Efficiency in factor markets.* If households and firms view factor price as fixed and maximize utility and profits respectively, factors will be hired where their marginal value product equals the subjective reservation value of the factors to their owners.
4. *Global efficiency.* Since equilibrium prices established in goods and factor markets ensure equality of buying and selling prices, there will be no excess demand for, or supply of, goods and resources. Furthermore, marginal rates of transformation in production will be equated with marginal rates of substitution in consumption.

Lange defines a very special type of socialist system.[5] Although the means of production are publicly owned, prices of consumer goods and labor services are established in markets. Thus there is presumed freedom of choice in consumption and occupation. Prices of capital goods and intermediate goods (produced goods not for final consumption, such as steel and other industrial materials) are fixed by a central planning board (CPB). Given these conditions, resource allocation is achieved in the following ways:

1. *Efficiency in consumption.* Since consumers purchase goods in markets, efficiency in consumption in the Lange model is achieved in the same way as in the market model. The same is true with the supply of labor.
2. *Efficiency in production.* Firm managers are no longer required to maximize profits. Instead, the CPB imposes the following rules:
   (a) Factors must be combined in such a way that the cost of production is minimized for every level of output. This is assured if the marginal rate of factor substitution is equal to the ratio of factor prices, which satisfies one of the Pareto criteria for efficiency.
   (b) Production should be expanded to the point where price

[5] Lange's model is described in full in *Economic Theory of Socialism.*

equals marginal cost. This ensures that goods will be produced at equal marginal rates of transformation throughout the economy.

The two conditions imposed on Lange's enterprise managers are equivalent to imposing the condition that they maximize profits when prices are fixed.[6] Presumably, Lange had an ideological aversion to requiring profit maximization on the part of managers, even though it may be easier to obtain compliance by using a profit incentive rather than requiring adherence to the Lange rules. However, the important condition of his model is that goods and factor prices are not set by firms, no matter how large the firm in relation to the market, but are viewed by them as fixed. In this case, either by Lange's two rules or by a profit-maximizing criterion, efficiency in production will be achieved.[7]

3. *Global efficiency*. Since prices of many goods are not established in markets, there is no assurance that the quantity supplied of any good will equal the quantity demanded. Thus, although there is efficiency in production (all firms producing at equal *MRT*) and in consumption (all households consuming at the same *MRS*), there may not be global efficiency if markets are not cleared at these rates.

Lange solves the problem by obliging the CPB to readjust prices continually in response to shortages and surpluses of the commodity in question. He states,

> Thus, the accounting prices in a socialist economy, far from being arbitrary, have quite the same objective character as the market prices in a regime of competition. Any mistake by the Central Planning Board in fixing prices would announce itself in a very objective way—by a physical shortage or surplus of the quantity of the commodity or resource in question—and would have to be corrected in order to keep production running smoothly.[8]

Lange argues that prices in competitive socialism are likely to equalize demand and supply faster than in a market system because the CPB can consider the interrelations of many markets at once and allow in advance for possible effects of a change in one price on other prices.

In essence, Lange views the allocative aspects of his model as nearly identical to those of a market model. He concludes,

> Our study of the determination of equilibrium prices in a socialist economy has shown that the process of price determination is quite analogous to that

[6] The assumption is that all prices in the economy will be viewed by participants as fixed. Prices of intermediate products and capital are set by the CPB. Lange assumed that consumer goods are produced and sold under competitive conditions, that is, that no one seller is large enough to influence market price.

[7] The Yugoslav economy, which uses a variation of the Lange scheme for allocating resources, relies on profit sharing, and hence a profit incentive, to ensure efficient use of resources.

[8] Lange, *Economic Theory of Socialism*, p. 82.

> in a competitive market. The Central Planning Board performs the functions of the market. It establishes rules for combining factors of production and choosing the scale of output of a plant, for determining the output of an industry, for the allocation of resources, and for the parametric use of prices in accounting. Finally, it fixes the prices so as to balance the quantity supplied and demanded of each commodity. It follows that a substitution of planning for the functions of the market is quite possible and workable.[9]

Although the formal principles of resource allocation are the same, however, the actual allocations achieved in capitalist and socialist states will generally be quite different. This is because of differences in policies with respect to income distribution, externalities, and growth. We will examine each of these in turn.

Income Distribution

The ideology of a socialist system provides a built-in value judgment concerning optimal income distribution. Therefore, socialists are explicitly concerned with measures that will reduce income inequality. However, there is still a concern that the redistributive mechanism should not interfere with efficient allocation of productive resources. In a socialist economy income distribution tends to be less unequal than in capitalism even before explicit redistributive measures are undertaken, since the only source of individual income is wages. All capital and land are socially owned. Furthermore, the returns to capital and land that accrue to the state and that are in excess of requirements for public goods are distributed to workers as a social dividend. Lange is careful to point out that the social dividend should be distributed in such a way so as not to interfere with occupational choice. He recommends that the social dividend be divided equally among all individuals or families. This method would reduce the inequality resulting from productivity differentials.

Lange also advocates progressive taxes on labor income as a partial means of financing public goods and for redistributional purposes. Inequality can also be reduced by collective provision of medical care, transportation, housing, and other goods and services consumed privately in most capitalist economies.

Externalities

Lange argues that the board could take externalities directly into account in establishing prices. However, as in a market system, it is generally difficult to place an actual value on external effects. If an externality takes the form of a public good such as air pollution, it may be difficult to determine the social cost and the consequent "tax" to be added to the price of the product of the polluting industry. This problem is inherent in designing effective antipollution policy in a capitalist economy as well, however, and the discretion of the CPB over prices

[9] Ibid., pp. 82–83.

in the Lange scheme would facilitate making such charges. In practice, then, the implementation of taxes or subsidies to industries generating externalities is easier under the Lange plan than in a market system. The problem of acquiring the information to ascertain the correct amount of tax or subsidy is common to both systems.

Growth

The rate of growth in the Lange model is not determined by consumers' preferences for present over future consumption but by the decision of the CPB to expropriate resources for investment. Since many actual socialist economies are at low or intermediate levels of economic development, a rapid rate of growth has generally been an overriding social goal. The CPB can encourage firms to increase investment by reducing the price of capital. Since it controls all nonwage sources of income, these can be made available for investment purposes at no apparent cost to society.[10]

4–5 EVALUATION OF THE LANGE MODEL

Critics of the Lange model have been quick to point out the difficulty for the CPB of finding a set of prices that would in fact clear all goods and factor markets simultaneously. Without actually solving the equations reflecting tastes and technology, the CPB must fix prices on an arbitrary basis. Although scarcity and surplus will cause the CPB to revise price schedules, the changes can take place only after the CPB has been notified and the new prices announced to all parties concerned. Branko Horvat notes that in Yugoslavia under administratively set prices there was chronic excess demand. Most consumer goods were rationed and sold at the existing uniform prices, which were too low to clear the market.[11]

Critics have also noted that price fixing would involve establishing individual prices for all varieties of goods. Horvat points out that in Yugoslavia sellers avoided CPB-established prices by making small changes in the design of the product, thus transforming it into a new product not subject to price control. He notes that in 1964 almost 25,000 such "new products" were launched.[12]

The Problem of Incentives

One of the most frequent concerns of critics of the Lange model has been the lack of incentives for managerial responsibility and initiative. How, in fact, can

[10] The actual cost, of course, is the utility value of the current consumption foregone by diversion of resources to investment.

[11] Branko Horvat, "Yugoslav Economic Policy in the Post-War Period," *American Economic Review, LXI* (June 1971), supplement, p. 109.

[12] Ibid., p. 111.

the CPB ensure that enterprise managers will follow the rules of marginal cost pricing and minimizing unit cost at each scale of output without some incentive based on profits? It has been argued that rewarding managers on the basis of enterprise profitability would produce the desired results if the firm views CPB-established prices as fixed.

Suppose managers were rewarded on the basis of profitability, however. Assuming the firm is large enough to affect market supply appreciably by its activities, would it not be to its advantage to restrict output in the same manner as a capitalist manager? By restricting supply and causing a shortage, the firm could force the CPB to raise the product price. This means that the firm's activities can affect market price and the maximum profit position would be where marginal revenue equals marginal cost, rather than where price equals marginal cost.

Institutional Considerations

In order to administer the complex price system and to ensure that firm managers are playing the game fairly, a huge bureaucracy would be required. The costs of acquiring information and administering the system would divert resources from other uses and hence reduce the efficiency of the economy in providing final goods and services. Some critics also suspect that the all-pervading economic power of such a bureaucracy would have undesirable effects, particularly in view of the fact that the CPB can determine income distribution as well as the rate at which resources are diverted from consumption into investment. Horvat, Dobrinčić, Kidrić, and other Yugoslav Marxists have criticized state ownership as being a "remnant of capitalism," leading to an "increase and strengthening of privileged bureaucracy as a social parasite." [13]

Nevertheless, as a system for allocating resources the competitive socialist model has certain advantages over the market model. In the next section we will compare the two systems in this context.

4-6 THE MARKET AND NONMARKET MODELS COMPARED

Given certain assumptions, a market system will produce efficient resource allocation with minimal costs of acquiring and processing information. Information is transmitted through the automatic mechanism of the price system. The model requires that

1. All households and firms be so small that the activities of any one have only a negligible impact in any market.
2. There are no externalities.

[13] Cited in Horvat, "Yugoslav Economic Policy," p. 76. Yugoslavia has attempted to avoid the problem of centralized control in a socialist state by the establishment of the concept of worker management. The cited article by Horvat is an excellent description of the system.

3. There is no need to provide public goods.
4. The market mechanism operates sufficiently well so that equilibrium is attained quickly.

It was noted that when these conditions are violated, a market system breaks down in providing efficient resource allocation. Furthermore, a market system distributes income on the basis of factor ownership and productivity, which may not be desirable from the viewpoint of society.

In evaluating a system, it is important to consider its flexibility in rectifying inefficiences. One of the most serious difficulties in a market system is to regulate the activities of enterprises that have grown large because of technological economies of scale. If they are broken into smaller units, they lose this technological efficiency. Yet it is difficult in the ideological context of most capitalist systems, which rests largely on the laissez faire tradition, to justify government regulation of private industry.

In a Lange model the problem is theoretically solved, since regardless of size, firms view prices as predetermined, that is, beyond their influence. There are problems with this, too, however. If managers are simply told to produce where price equals marginal cost, there is no incentive for them to obey the rule and consequently there is an enforcement problem. If they are given an incentive to maximize profits, however, they can "cheat" by restricting supply and causing shortages, thus forcing the CPB to raise prices. In that case, they no longer take prices as given and behave essentially like capitalist monopolists.

Externalities provide a problem for both the market model and the competitive socialist model. In both cases, information is required concerning the extent of the interaction effects of the externality and the social value (or cost) of the externally generated output. Given the authority vested in the CPB over the price system as well as its experience in dealing with administered prices, it is likely that external effects could be accounted for and the correct social adjustments made in a Lange price system better than in a market model. In a market model adjustment for externalities can theoretically be introduced in a system of taxes and subsidies. However, the likelihood of developing the mechanism for universal application in actual practice is low.

Any economy must deal with the need for providing public goods. As has been mentioned, a price system of any sort—be it the market or Lange variety—is not appropriate for financing public goods. Consequently, taxation or other means of financing goods must be worked out outside the price system in both models.[14] Both models provide ample allowance for taxation for this purpose. However, a socialist country, because of its egalitarian ideology, has less difficulty appropriating taxes for public projects (collective consumption) than a capitalist state. In the latter, hospitals, transportation facilities, and housing

[14] We saw that there may be some justification on efficiency grounds for taxing the rich at a higher rate than the poor for public goods. This principle applies to both systems, however, and is generally practiced on distributional grounds anyway.

are more likely to be privately financed. This doesn't necessarily interfere with efficiency per se, since it relates to differences in social preference, that is, what is to be produced rather than the efficiency in providing it.

Lange's arguments to the contrary, it is likely that the market mechanism is a faster and less costly means of achieving balance between supply and demand in goods and factor markets. Alfred Whitehead once remarked, "Civilization advances by extending the number of important operations which we can perform without thinking about them."[15] Observation of centrally planned economies today makes it abundantly clear that a practical solution to the problem of balancing demand and supply under a system of administered prices is one of the most seriously deficient aspects of the system.

Finally, with respect to income distribution, the CPB has the ability to effect desired changes that are not forthcoming in a market system. It is important to remember, however, that this can only be viewed as a virtue to the extent that there is general agreement in the society concerning what the optimal income distribution should be. In socialist states the Lange model is viewed as preferable to the market model for this reason. However, these socialist societies are committed to an ideology that predefines the optimal income distribution as egalitarian. If such agreement does not exist, the CPB could become the tool of the dominant political power and would not necessarily act in the interests of all members of society. In other words, the concept of a social optimum presupposes some unanimity of objectives among the individuals who comprise the society. Unanimity concerning distributional issues is more likely to be present in a socialist than in a capitalist state.

In capitalism, however, there is a general acceptance of the desirability of individual freedom, a question the socialists tend to sidestep.[16] The market model and the model of competitive socialism could both, in theory at least, provide efficient resource allocation. Consequently, the choice of system will ultimately depend on its consistency with the noneconomic goals of the society and on the likelihood that it can achieve the desired resource allocation in actual practice.

4–7 NONMARKET DECISIONS IN A MARKET ECONOMY

We have observed that in an economy that relies on the market mechanism for allocating most of its goods and resources there will always be a need to make some allocative decisions outside of markets. The most notable example is the public sector, but such decisions must also be made within large corporations. Furthermore, some central control of resource allocation where important externalities exist may be deemed desirable. Can the Lange model of competitive

[15] Cited in Hayek, "The Use of Knowledge in Society," p. 528.

[16] On this point see Milton Friedman, *Capitalism and Freedom* (Chicago: University of Chicago Press, 1962), p. 19.

socialism provide insights into optimal nonmarket planning techniques for a market economy?

Suppose we are interested in designing a policy that will prohibit emission of sulfur oxides beyond a certain level. One way to do this would be to identify all industries that produce sulfur oxides as a by-product, compute the rate of emission of each firm as a function of the output of each firm, and then find the best way to restrict sulfur oxide emission, that is, restrict emission by firms in such a way so as to cause the least reduction in output. Clearly, this would involve different restrictions for each firm, since each firm will generally emit sulfur oxide at a different rate per unit of output. Thus the cost of pollution control in terms of foregone output will be different for different firms.[17] Finally, each firm must be advised of the restriction applicable to it. Restricting sulfur oxide emission should cause firms to switch to production processes that emit less sulfur oxide. Then the whole quota system would have to be revised.

An alternative, consistent with the Lange framework, would be to establish a tax on emitted sulfur oxides equal to the estimated social cost of the resultant pollution.[18] Taxes could vary geographically to the extent that social costs of pollution vary, for instance, between Newark and the Mojave Desert. If the effect of the tax were to reduce sulfur oxide emission too little or too much, the tax could be raised or lowered.

Such a tax would involve the acquisition of considerably less information than the quota system. Furthermore, enforcement would be limited to tax collection—a technique that is well developed in the United States—rather than visits to individual factories to check rates of sulfur oxide emission. The tax would cause firms to internalize the social cost of pollution so that all costs of production (including the social cost of pollution) would be taken into account in the firms' calculation of how and how much of the product should be produced. Product prices would reflect true social marginal cost.

In addition to lower administrative and enforcement costs, effluent charges (taxes on pollution) are generally preferred to direct controls on pollution on other grounds. There would be no incentive to "cheat" by under-reporting sulfur oxide emissions to the control authority. More important, however, in a market-

[17] If the social cost of pollution is the same in all geographic areas (a doubtful assumption), at the optimum firms will all emit sulfur oxide at the same rate per unit of output. Otherwise, emission may occur at different rates.

[18] The tax could also be levied on the basis of output if the aggregate rate of sulfur oxide emission were known. In this case the tax would be set equal to the marginal rate of transformation between sulfur oxide and the output, multiplied by the estimated social cost of the resultant pollution, that is,

$$t = p_S \frac{\Delta S}{\Delta X}$$

where p_S is the social cost of sulfur oxide emission and $\Delta S/\Delta X$ is the marginal rate of transformation between sulfur oxide and output. This assumes that all firms are producing sulfur oxide at the same rate per unit of output. If emissions cannot be accurately measured, an excise tax would be a second best solution.

oriented society, the use of a price system to guide individual choice is not viewed as a threat to individual freedom. Use of a quota system to guide resource allocation may be viewed as inconsistent with the ideology on which a market economy is based.

QUESTIONS FOR STUDY AND REVIEW

1. "If efficient resource allocation is the goal of an economy, then prices must exist in that economy." Comment.
2. What is the theoretical justification for Lange's model of competitive socialism? Under what circumstances does it compare favorably or unfavorably with the market model of laissez faire capitalism?
3. Discuss the desirability of the use of output restrictions as compared with effluent charges in reducing the level of air pollution. Be sure to consider questions of allocative efficiency as well as enforcement.
4. Compare and contrast the efficacy of a market system with that of a nonmarket system for producing a socially optimal allocation of resources.

ADDITIONAL READING

Dobb, Maurice. *Welfare Economics and the Economics of Socialism*. Cambridge, England: Cambridge University Press, 1969, chapter 7.

Hayek, Friedrich A. "Socialist Calculation: The Competitive 'Solution,'" *Economica, VII* (May 1940), pp. 125–149.

______. "The Use of Knowledge in Society," *American Economic Review, XXXV* (September 1945), pp. 519–530.

Koopmans, Tjalling C. "Efficient Allocation of Resources," *Econometrica, XIX* (October 1951), pp. 455–465.

Lange, Oskar. *On the Economic Theory of Socialism*. Minneapolis: University of Minnesota Press, 1938.

Schumpeter, Joseph A. "The Nature and Necessity of a Price System," in *Economic Reconstruction*. New York: Columbia University Press, 1934, pp. 170–176. Reprinted in David R. Kamerschen, ed. *Readings in Microeconomics*. New York: World Publishing Co., 1967.

part II
TOOLS OF ANALYSIS

5 The Household Sector

In the previous chapters we discussed the goals and rationale for a theory of resource allocation. In the chapters that follow we will develop the analytical tools necessary for evaluating the allocative mechanism under different circumstances.

Since the traditional criterion for evaluating economic performance is the satisfaction of the wants of the people in society, it is useful to begin our discussion with a model of the household sector. Such a model should be able to predict how consumers will react to changes in economic conditions such as changes in taxes, in their income, or in the prices of the various goods and services they buy. The model can also be used to establish the conditions under which the efficiency criteria are met. In addition, some economists have attempted to derive insights into social preferences on the basis of models of individual behavior.

5-1 THE CLASSICAL APPROACH

The earliest formal model of the household sector was developed in the latter part of the nineteenth century and was popularized by Alfred Marshall in his *Principles of Economics,* which appeared in 1890. The classical analysis was based upon a model of household tastes that viewed the consumer as an abstract psychological entity. This theoretical consumer experiences psychic gratification from goods and services, called **utility.** Despite the Benthamite label, utility is not a measure of usefulness or need but a measure of the desirability of a commodity from the psychological viewpoint of the consumer. For any consumer, utility is measurable and one-dimensional in the sense that it can be compared among all commodities consumed. Thus utility experienced from eating lunch can be compared with utility experienced when a new car is purchased.[1] The classical economists called this property of comparability **cardinality.**

---

[1] In the earliest versions it was assumed that utilities of various goods were independent and additive, that is, that the utility experienced from drinking coffee is unaffected by how much tea you have just drunk. But it soon became evident that this assumption was not only unrealistic but unnecessary, and it was dropped.

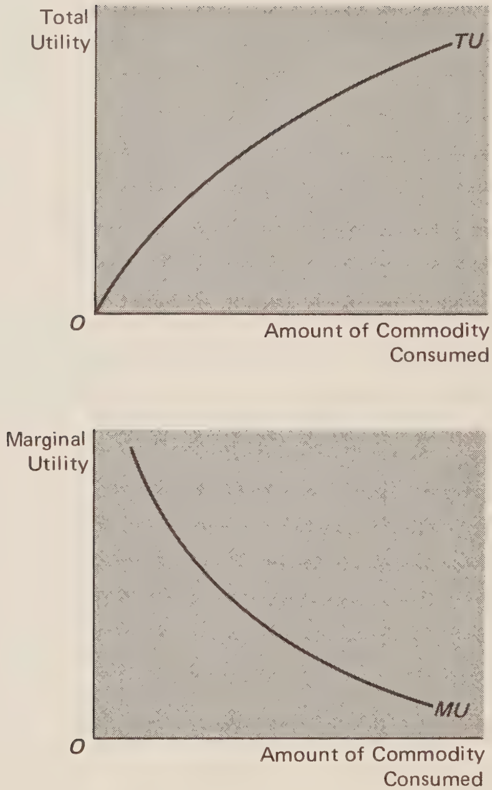


FIGURE 5–1
The Classical Utility Function

Cardinal utility is also comparable among individuals. This implies that the psychic satisfaction I receive from consuming goods and services can be measured and compared with that which you experience when consuming goods and services. As we shall see, the classical economists felt that this assumption was necessary to draw conclusions about the impact on social welfare of changes in the distribution of income.

The relation between utility and the amount of goods and services consumed by a household is called a **utility function.** Algebraically, the utility function can be written

$$U = f(x_1, x_2, \ldots, x_n)$$

where U is utility and x_i is the amount of the ith good or service consumed.[2] The classical utility function is shown in Figure 5–1. Since more will always be preferred to less, utility must rise as more of a commodity is consumed. But the principle of saturation suggests that the more consumed, the smaller the incre-

[2] If utility were additive, the utility function would be written

$$U = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

ment to total utility of successive units. Thus, although total utility increases with consumption of a good, marginal utility, the change in total utility associated with the consumption of an additional unit, decreases.

Substitutes and Complements

If the utility derived from a good depends on how much of it is consumed, it follows that utility also depends on the amount of certain other goods consumed. The utility of a cup of coffee obviously depends on the amount of tea you have just consumed; the utility of a tennis ball is clearly dependent upon the possession of a tennis racket.

Now, if the consumption of a good, *Y*, reduces the marginal utility of another good, *X*, then *X* and *Y* are said to be **substitutes.** Coffee and tea are substitutes. The relation between the marginal utility of a fixed amount of coffee and the amount of tea consumed is shown in Figure 5-2.

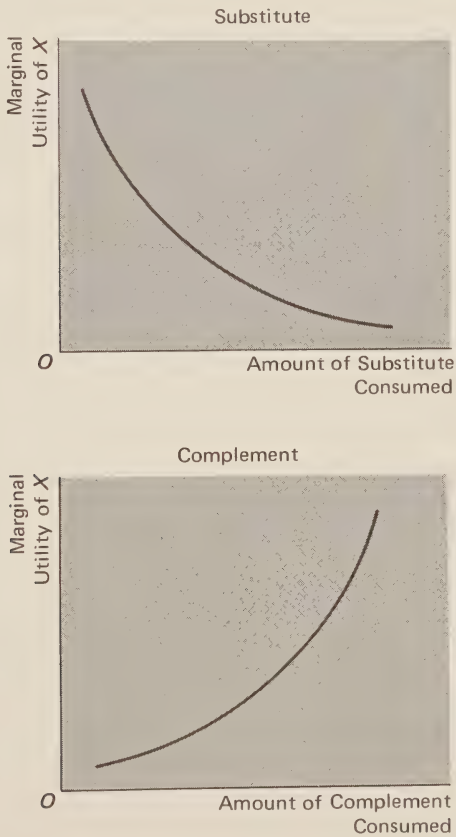


FIGURE 5-2
Effect of Substitutes and Complements on the Marginal Utility of a Fixed Amount of Commodity *X*

If the possession of Y increases the marginal utility of X, as shown in Figure 5–2, the goods are said to be **complements.** A tennis racket is complementary to a tennis ball.

If the possession of Y has no effect on the marginal utility of X, the goods are said to be **independent.** The possession of a tennis racket has a negligible effect on the marginal utility of a cup of coffee.

Recently, economists have observed that the utility functions of different individuals may be interrelated. The utility I experience from the purchase of a new Volkswagen may be considerably diminished if my neighbor drives up in a new Cadillac. On the other hand, I may receive psychic satisfaction from my friend's good fortune.

Theory of Consumer Behavior

Given a limited income, the consumer will distribute his purchases in such a way as to maximize his total utility. When total utility is maximized, the consumer is said to be in **equilibrium,** that is, he will not readjust his purchases unless the underlying conditions, prices, income, tastes, and so on, change. Since marginal utility declines the more of a good is purchased, the marginal utility per dollar spent on each good should be equal for all goods. Algebraically, the consumer maximizes utility where

$$\frac{MU_X}{p_X} = \frac{MU_Y}{p_Y}$$

MU is marginal utility, and p_X and p_Y are the prices of the goods X and Y. Marshall called the marginal utility per dollar spent on each good the **marginal utility of money,** $MU_\$$. In contemporary analysis $MU_\$$ is often called the **marginal utility of income.** For any single good, the ratio of the marginal utility to its price is equal to the marginal utility of income when the consumer is maximizing utility, that is,

$$\frac{MU_X}{p_X} = \frac{MU_Y}{p_Y} = MU_\$$$

The Demand Curve

Economists are often interested in how a change in the price of a good will affect consumer purchases. Suppose we wish to relieve traffic congestion on a bridge by levying a toll. If we wish to reduce the flow of traffic by 500 cars per hour, how much should we charge per automobile?

The relation between the price of a good and the quantity demanded is called the **demand function.** The demand function for a good of a household is called an individual demand function. The demand function for the good of all households taken together is the market demand function.

From the previous equation, the price a consumer is willing to pay depends on the marginal utility of the good in relation to the marginal utility of income in general, that is,

$$p_X = \frac{MU_X}{MU_\$}$$

Marshall assumed that for most goods, changes in the price of *X* or the quantity of *X* consumed have a negligible impact on the marginal utility of income. Therefore, $MU_\$$ can be treated as a constant, λ, and price will be proportional to marginal utility, that is,

$$p_X = \frac{1}{\lambda} \times MU_X$$

Thus the marginal utility schedule for the good has the same shape as the individual demand function. A demand curve is the graphical representation of the demand function. Since marginal utility is always decreasing, the demand curve for a good must be negatively sloped.

If the good is very important in the consumer's budget and if the price change is very great, an increase in the price will have the effect of reducing the real income, or purchasing power, of the consumer. This will increase the marginal utility of money. For such goods, the demand curve will be flatter than the marginal utility curve for the good as shown in Figure 5-3. In extreme cases, the **income effect** will cause the demand curve to be positively sloped. If a price increase causes more to be consumed, the good is called a **Giffen good.**[3]

A recent study has shown that when food prices are rising, the consumption of beans, eggs, and starches increases while that of meat declines, even when meat prices are rising more slowly than other food prices. Apparently, the income effect, that is, the decline in purchasing power associated with higher food prices, is so great for many people that they can no longer afford meat and must switch to cheaper foods, even though their prices are rising faster than meat prices. Beans, eggs, and starches are **inferior goods;** the demand increases when real income, or purchasing power, falls. They are also Giffen goods if the demand increases as their prices rise.

Because most goods represent a small part of the consumer's budget, we can assume that the income effect will be negligible and that the demand curve will be negatively sloped. Thus an increase in price will cause a reduction in quantity demanded. However, it is important to recognize that in cases where the commodity is a large part of the consumer's budget, the impact of a price increase will be more difficult to predict.

[3] Giffen's paradox refers to goods for which the quantity demanded rises with price. The paradox was first noted by Sir Francis Giffen in connection with the Irish potato famine of 1845. As the scarcity of potatoes raised potato prices, poor families suffered a reduction in real income and were forced to reduce consumption of meat and other more expensive foods. To compensate, they increased consumption of potatoes, which were relatively cheaper, despite the increase in potato prices.

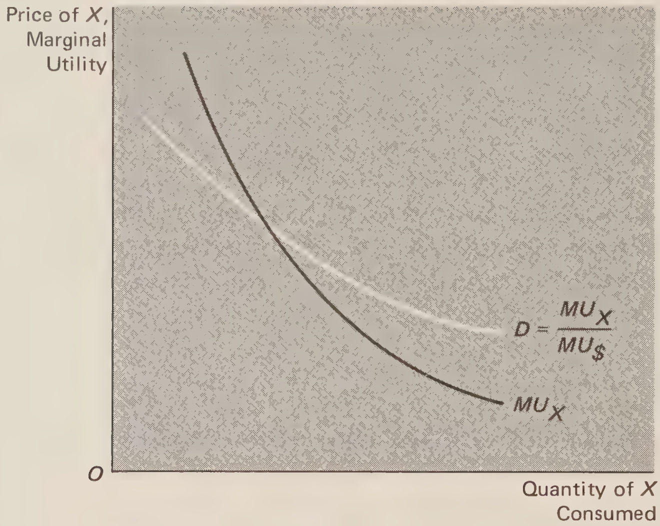


FIGURE 5–3
Marginal Utility and the Demand Curve

The Market Demand Curve

Policy makers and firms are usually interested in determining how consumers behave as a group. Since total market demand is the sum of the purchases of all households, the market demand curve is the horizontal sum of the individual demand curves.[4] Suppose only two individuals purchase commodity X. The derivation of the market demand curve is shown in Figure 5–4. When price is p_1, household I consumes OA and household II consumes OB. The market demand at p_1 is OC, which is the sum of OA and OB, that is,

$$OC = OA + OB$$

At p_2 the market demand is OF, the sum of OD and OE. Notice that the change in market demand with respect to price is the sum of the individual changes, that is,

$$CF = AD + BE$$

If all the individual demand curves are negatively sloped, then the market demand curve will be negatively sloped. However, a good may be an inferior good for some households and yet market demand may still decline when the

[4] This is true only for a pure private good. For a public good, the total social consumption is the same as each individual's consumption. For such a good, the market demand curve is the *vertical* summation of individual demand curves. For a discussion see pages 50–53 in this book; Paul A. Samuelson, "Diagrammatic Exposition of a Theory of Public Expenditure," *Review of Economics and Statistics, XXXVII* (November 1955), pp. 350–356; and Richard A. Musgrave, *The Theory of Public Finance* (New York: McGraw-Hill, 1959), chapter 4.

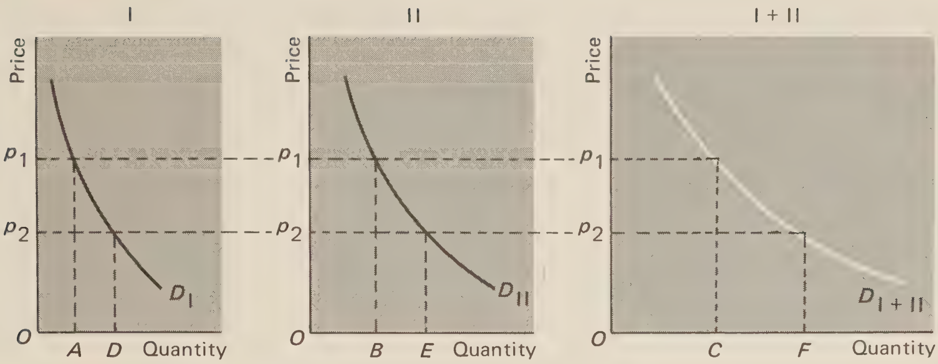


FIGURE 5-4
Derivation of the Market Demand Curve from Individual Demand Curves

prices rise. This depends on the relative importance of those households in total market demand. It is likely that for poor families there are many inferior goods. When budgets are small, no item has a negligible impact on purchasing power. A rise in clothing prices may cause these families to buy more sweaters instead of a coat. But poor families represent a small part of the total demand for sweaters and other goods that are inferior for them. Therefore, the market demand curve for such goods is likely to be negatively sloped. Only goods that are inferior for an income-weighted majority of families will be inferior to the market as a whole.

Characteristics of Market Demand Curves

Let us return to our problem of designing a toll to reduce traffic flow. To determine the toll required to reduce the flow of traffic by 500 cars per hour, we would need to know the slope of the market demand curve, that is, the ratio of vertical to horizontal change, from the initial point of consumer equilibrium to the desired point. In this case, initial equilibrium occurs where p is zero, or where the demand curve intersects the horizontal axis, and the desired point is where quantity q is reduced by 500.

A hypothetical market demand curve for the use of the bridge is shown in Figure 5-5. When no toll is levied, the flow of traffic is 1,000 cars per hour. To reduce the traffic flow by 500 cars, a toll of 50 cents must be levied on each car. The slope of a line connecting those points on the demand curve is

$$\frac{\Delta p}{\Delta q} = \frac{50}{500} = \frac{1}{10}$$

To reduce the rate of traffic flow by 10 cars per hour, 1 cent must be added to the toll.

Suppose the traffic authority is interested not only in relieving traffic congestion but also in raising revenue to build an additional bridge. If the traffic

authority were to levy a toll of 50 cents, the rate of traffic flow would be 500 automobiles an hour and total hourly receipts from tolls would be $250. From the demand curve in Figure 5–5, if the toll were raised to $1.00, traffic flow would be reduced to 300 cars per hour but receipts would rise to $300.

It would appear that the higher the toll the greater would be the receipts. But if the toll were too high, the reduction in the traffic flow would more than offset the effect of the increased toll on total receipts. If the toll were raised to $2.00, the hourly traffic flow would fall to 100 and total receipts from tolls would decline to $200. There must be some amount of toll consistent with maximizing receipts.

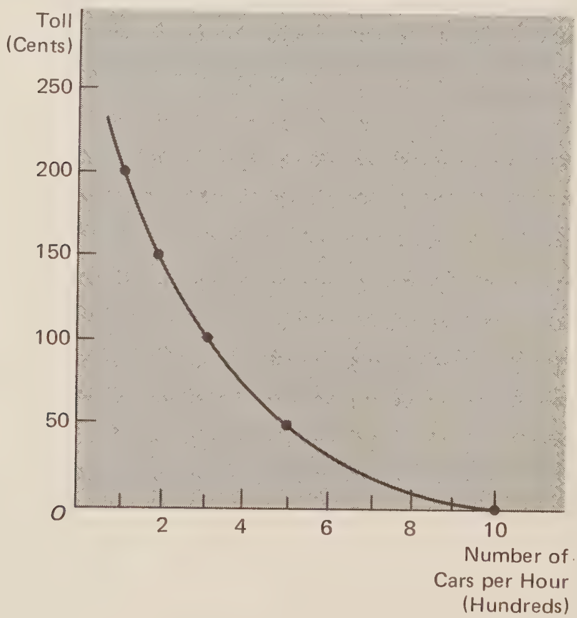


FIGURE 5–5
Market Demand Curve for Use of a Bridge

Since the effect of a change in the toll on total receipts clearly depends on the sensitivity of the traffic flow to a change in the rate, it may appear that the slope of the demand curve would be an important consideration. But the amount of the toll is measured in monetary units and the traffic flow in number of automobiles per hour. We could arbitrarily change the units of measure (from cents to dollars or cars per hour to cars per minute) and change the slope of the demand curve. Clearly, if the slope is 1/10, it makes considerable difference to total revenue collected whether a 1-cent increase in the toll reduces traffic flow by 10 cars an hour or by 10 cars a minute.

Consequently, economists use a different measure of the sensitivity of demand

to changes in price: **the elasticity of demand.** Elasticity of demand relates per cent change in quantity demanded to per cent change in price. Algebraically,

$$E_D = -\frac{\Delta q/q}{\Delta p/p} = -\frac{p}{q} \times \frac{\Delta q}{\Delta p}$$

Since the slope of the demand curve is negative, $\Delta q/\Delta p$ is also negative. We conventionally put a minus sign before the expression so that we can refer to demand elasticity in positive terms.

Suppose we want to determine if a reduction in the toll will increase total revenue. Quantity demanded will increase as the toll is reduced. But if E_D is less than 1, the per cent increase in quantity demanded will be less than the per cent decrease in the toll and total receipts will decline when the toll is reduced. If E_D equals 1, the per cent increase in quantity demanded will be exactly equal to the per cent decrease in price and there will be no effect on total receipts. On the other hand, if E_D is greater than 1, the per cent increase in quantity demanded will be greater than the per cent decrease in price and total revenue will rise when the toll is reduced.

The change in total receipts when quantity demanded increases by an additional unit is called **marginal revenue.** If marginal revenue is positive, an increase in quantity demanded will increase total receipts. Thus a reduction in price when marginal revenue is positive will cause total receipts to rise. A positive marginal revenue is always associated with a demand elasticity greater than 1. When marginal revenue is zero, demand elasticity is 1, and total receipts are maximized.

Table 5–1 shows marginal revenue, *MR,* and demand elasticity, E_D, for the

TABLE 5–1
Demand Schedule for Use of a Bridge

Amount of Toll (Cents)	Number of Cars per Hour	Total Receipts (Dollars)	Change in Total Receipts	Elasticity of Demand
p	q	$TR = pq$	$MR = \frac{\Delta TR}{\Delta q}$	$E_D = -\frac{p}{q} \times \frac{\Delta q}{\Delta p}$
200	100	200		
			100	$2\frac{3}{4}$
150	200	300		
			0	1
100	300	300		
			−50	$\frac{13}{15}$
50	500	250		
			−250	undefined *
0	1,000	0		

* Would require division by zero.

demand curve for the use of the bridge. The relation is shown graphically in Figure 5–6. If the toll is above \$1.50, marginal revenue is positive and demand elasticity is greater than 1. A reduction in the toll will increase total receipts. MR and E_D both decline as the toll is lowered.[5] For tolls below \$1.00, MR is

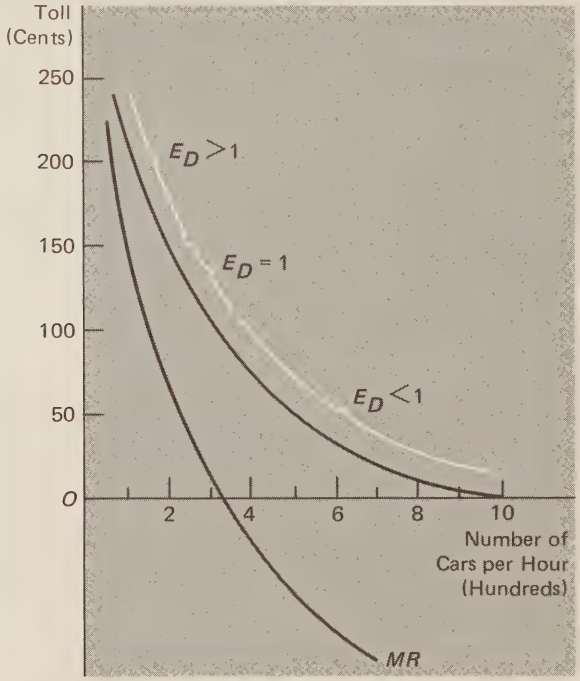


FIGURE 5–6
Demand Curve and Marginal Revenue for Use of a Bridge

negative and E_D is less than 1. A reduction in the toll below \$1.00 will cause total receipts to decline. If the traffic authority wishes to maximize total receipts, it will levy a toll between \$1.00 and \$1.50. The revenue-maximizing toll is found in the range of the demand curve where elasticity is equal to 1 and marginal revenue is zero.

[5] If the demand curve were linear, it would be easy to see why elasticity must decline as one moves down the demand curve. For a linear demand curve, the slope, $\Delta p/\Delta q$, is constant. Therefore, $\Delta q/\Delta p$, is also constant. Since the quantity p/q declines as one moves away from the vertical axis, the elasticity,

$$E_D = -\frac{p}{q} \times \frac{\Delta q}{\Delta p}$$

must also decline.

Welfare Considerations

We know that the revenue-maximizing toll lies between $1.00 and $1.50. What other considerations must be taken into account? Clearly, the amount of the toll will affect traffic flow on the bridge. If the toll is $1.50, only 200 cars will cross the bridge per hour. If it is $1.00, 300 cars will use the bridge.

Consider the demand curve for the use of the bridge shown in Figure 5–7. We may assume that the use of the bridge is a minor component of the consumers' budgets, so that the demand curves represent the sum of the marginal utility curves of all the consumers who use the bridge. (Hence, we ignore the income effect of a change in the toll.) A moment's reflection will verify that the

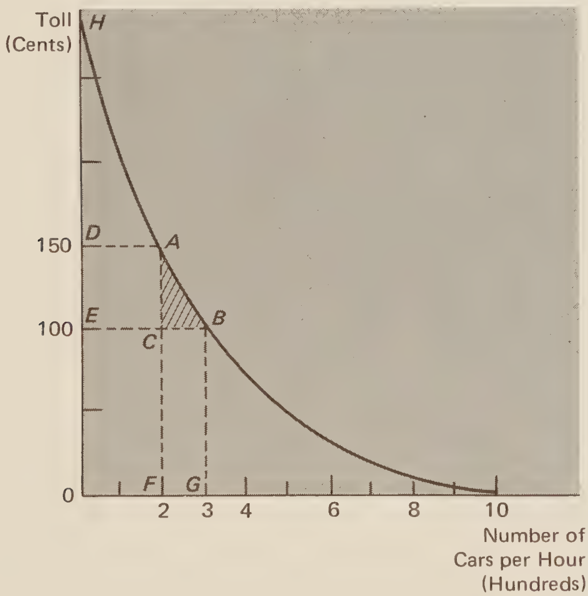


FIGURE 5–7
Impact of Alternative Tolls on Consumer Surplus

area under the marginal utility curve is total utility. Since *MU* is the change in total utility associated with consuming an additional unit, the area under the *MU* curve represents the sum of the marginal utilities for each unit, or total utility. Remember that the classical economists viewed utility as a measurable entity that could be compared among individuals. Thus the area under the *market* demand curve (the sum of the individuals' demand curves) represents the sum of all the individuals' total utility, or **social utility,** derived from using the bridge.

If the toll is $1.00, use of the bridge is *OG* and social utility is represented by the area *OGBH*. If the toll is $1.50, use of the bridge is *OF* and social utility is

OFAH. Social welfare is thus increased by the amount *AFGB* when the toll is reduced from \$1.50 to \$1.00, even though the amount consumers spend in tolls (and the amount collected by the traffic authorities) is unchanged. From the viewpoint of society, the lower toll is preferable, since it is associated with a higher level of social welfare.

Marshall noted that the amount consumers actually pay for a good is less than the amount they are willing to pay. Consumption of a good increases as prices fall, because individuals consume more of a good or because more individuals come into the market. (In the case of the bridge, increased hourly consumption will be due to more individuals coming into the market.) When the price falls, everyone benefits from the lower price, despite the fact that some people *would* have paid more. The difference between the price consumers would be willing to pay and the amount they actually pay is called **consumer surplus.** Consumer surplus represents utility or psychic satisfaction experienced by individuals that is not reflected in the amount they actually pay for the good.

In Figure 5–7, when the toll is \$1.50, use of the bridge is *OF* and consumer surplus is represented by the area *DAH*. When the toll is lowered to \$1.00, use of the bridge increases to *OG* and consumer surplus rises to *EBH*. The welfare gain to consumers of the reduction in the toll is the increase in utility associated with increased use of the bridge. This is represented by the amount the additional 100 cars per hour actually pay, *CBGF, plus* the gain in consumer surplus, the shaded area *CBA*. Although the gain in consumer surplus is not reflected in the amount paid by the additional users of the bridge, it must be taken into account when assessing the welfare effect of a reduction in the toll.

The assumption that utility is cardinal—that different people have comparable degrees of psychic gratification when they consume goods and that utility can be measured—was important for deriving social welfare propositions from the classical approach. If we cannot add up people's utilities, we cannot say unequivocally that a change is an improvement as long as anyone is harmed.

In this particular problem, however, the classical assumption may provide a useful first approximation. Even though we have no "utility meters" to assess the psychic pleasure of bridge users, we know that lower tolls allow more people to use the bridge. The distributive impact would favor the poor without harming the rich. And there is no cost to the traffic authority, since total collections will not fall.

But there is a final consideration in designing the optimal toll. Traffic congestion produces negative externalities, diseconomies, such as reducing the rate of speed, air pollution, and an increased accident rate. We know that a price system, which reflects private valuations of benefit and cost, does not take externalities into account. Users of the bridge are imposing a social cost on other users, reducing their marginal utility. A toll should take these social costs into account.

This is not a problem in this case as long as the estimated social cost of the use of the bridge is less than the toll. Recall that at the outset the traffic authority wished to reduce the hourly rate of traffic flow to 500. This was presumably

based on the idea that in the absence of a toll reflecting social cost the bridge was overutilized. Since the proposed toll of $1.00 reduces traffic flow to 300 per hour, it has reduced demand by more than that required for allocative efficiency. From an allocative point of view, therefore, the bridge is being underutilized.[6] Our "optimal toll" of $1.00 is not necessarily optimal in an allocative sense but is the toll that maximizes social welfare when the traffic authority is maximizing collections.

5–2 THE INDIFFERENCE ANALYSIS

Many economists were unhappy with the classical model. We have no way of actually measuring utility, and the assumption that psychic gratification is comparable among individuals or even among different goods for the same individual flies in the face of common sense. Even though the classical model was useful for explaining behavior of households, the implausibility of its assumptions made many people skeptical.

The indifference analysis abandons the notion of cardinal utility altogether. Instead, it simply assumes that each consumer would be able to tell us, if we describe any two alternative situations to him, whether he prefers one or the other or whether he considers them both equally satisfactory.

The basic tool of indifference analysis is the indifference map. As shown in Figure 5–8, quantities of the goods consumed are measured along the axes. The curves, U_1, U_2 and U_3 are called **indifference curves.** Along any indifference curve, U_1 for instance, the consumer finds various alternative combinations of X and Y equally satisfactory.

If the consumer could acquire more X without losing any Y, he would be better off. Consequently, along the indifference curve the consumer must always sacrifice some X to obtain Y. This implies that indifference curves are always negatively sloped.

Consider the slope of U_1 at point A. The slope represents the amount of Y that the consumer is willing to sacrifice to obtain an additional unit of X. This is the **marginal rate of substitution** of Y for X, MRS_{YX}. It is assumed that the more X the consumer has, the less willing he will be to sacrifice Y to obtain additional X. Therefore, MRS_{YX} decreases as one moves along the horizontal axis, making the indifference curve convex to the origin.

Points (combinations of Y and X) on U_1 are preferred to points below it. Points above U_1 are preferred to points on it. U_2 is an indifference curve that represents a higher level of satisfaction or utility. Notice that by utility we simply mean a state of well-being in a relative, or ordinal, sense, not a cardinal measure of psychic gratification. Combinations of Y and X on U_2 are equally satisfactory, and all of them are preferred to combinations on U_1. U_2 is also convex, since we

---

[6] We have assumed there is no cost to using the bridge other than the social cost of traffic congestion.

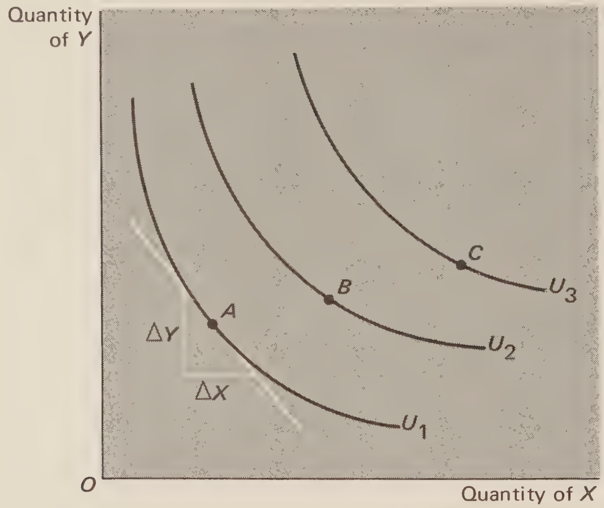


FIGURE 5–8
Indifference Map for a Single Consumer

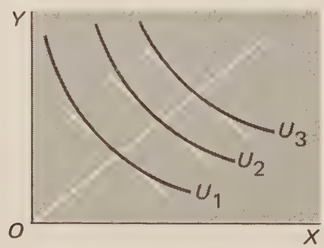
assume the MRS_{YX} declines as *X* is substituted for *Y*, regardless of the level of satisfaction.

Notice that the indifference curves do not intersect. This reflects the assumption that the consumer's preferences are **transitive.** If *C* is preferred to *B*, and *B* to *A*, then *A* cannot be preferred to *C*. If *A* were preferred to *C*, then U_1 would have to intersect U_3.

Although indifference curves are assumed to be nonintersecting, they are not necessarily parallel. If indifference curves were parallel with respect to the origin, this would imply that the relation between the MRS_{YX} and the relative amounts of *Y* and *X* consumed is independent of the absolute amounts of *Y* and *X* consumed.[7] Although this may be the case for some goods, it is not necessarily true for all.

Consider the indifference map for food and luxuries, shown in Figure 5–9.

[7] If the indifference curves are parallel, they have the same slope (MRS_{YX}) where they cross a straight line from the origin, as shown here. On this line the relative amounts of *Y* and *X* consumed remain constant.



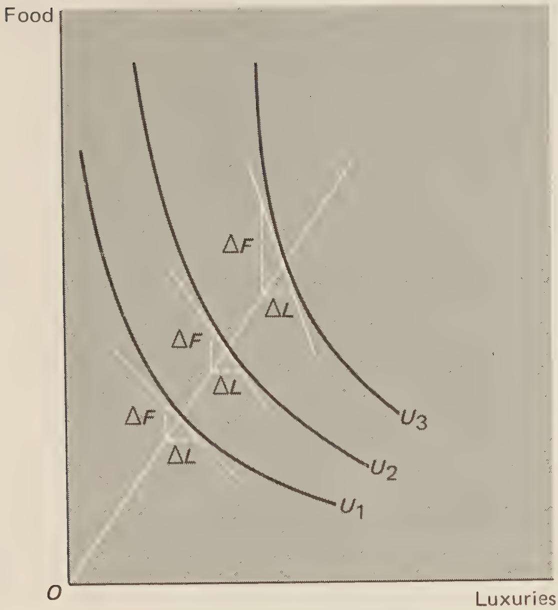


FIGURE 5-9
Indifference Map for Food and Luxuries

When food consumption is low, the consumer is not willing to exchange much food for luxuries and the MRS_{FL} is low. But as the consumer moves from U_1 to U_2 to U_3, the need for food is less urgent and the MRS_{FL} rises. Thus, although along *each* indifference curve the MRS_{FL} falls the more luxuries are consumed, as the consumer moves to higher indifference curves (higher levels of satisfaction) the MRS_{FL} rises when the *relative* amounts of food and luxuries consumed is the same. In this example, the **income elasticity of demand** (the per cent change in the quantity demanded relative to the per cent change in income) for food is low relative to that for luxuries. When two goods have different income elasticities of demand, the indifference curves will not be parallel with respect to the origin.

While the relation between indifference curves depends upon the relative income elasticities of demand for the goods, the shape of the individual indifference curves depends on how the marginal rate of substitution changes when more of one good is consumed. If goods are very close substitutes, such as beer and ale, *MRS* is unlikely to be affected very much by the relative amounts of the two that are consumed. An indifference map for beer and ale is shown in Figure 5-10. On the other hand, if goods are complementary, *MRS* will decrease rapidly as one is traded for the other. In the case of perfect complements, scotch and soda, for instance, *MRS* may be undefined, as shown in Figure 5-11. That is, there is no amount of scotch that could be traded for an additional soda that would leave the consumer equally well off.

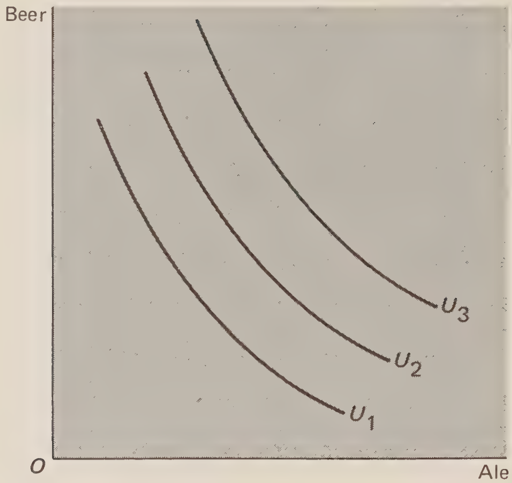


FIGURE 5–10
Indifference Map for Beer and Ale

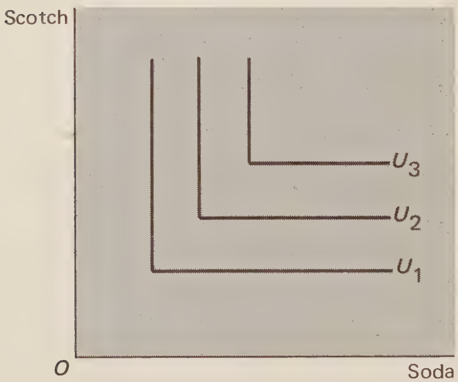


FIGURE 5–11
Indifference Map for Scotch and Soda

Consumer Equilibrium

The indifference map is a representation of the consumer's tastes or preferences. He must select that preferred bundle of commodities obtainable within his limited income. Obviously, the commodity bundles he can afford depend upon the amount of his money income as well as upon the prices of the various products.

Suppose the consumer allocates his budget of Y dollars between food, F, and luxuries, L, with prices p_F and p_L. Total expenditures can be represented algebraically as

$$Y = p_F F + p_L L$$

the alternatives available for Y dollars are shown along a **budget line** in Figure 5-12. Food and luxuries can be traded at a rate equal to p_L/p_F, that is, the consumer can obtain luxuries for food at that rate. Thus the slope of the budget line must be the negative of p_L/p_F.[8]

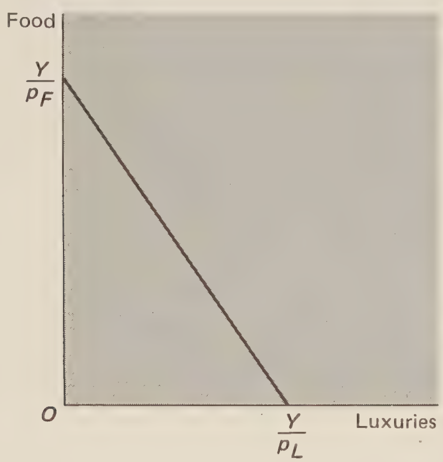


FIGURE 5-12
Food and Luxuries Obtainable for *Y* Dollars

Now compare the alternatives available to the consumer within his limited budget to his preferences as seen in his indifference map. In Figure 5-13 we have superimposed the budget line on the indifference map for food and luxuries. Point E on U_2 is the highest level of satisfaction attainable on the budget line. Since indifference curves are convex to the origin, the optimal position will be at the point of tangency between the budget line and the highest indifference curve, in this case, U_2.

Notice that since the budget line must be tangent to an indifference curve when the consumer is in his maximum utility equilibrium, their slopes must be equal there. Thus in equilibrium the condition

$$MRS_{FL} = -\frac{\Delta F}{\Delta L} = \frac{p_L}{p_F}$$

must hold. That is, the marginal rate of substitution of food for luxuries must be equal to the ratio of the price of luxuries to the price of food.

---

[8] The proof is as follows. Since Y, p_F, and p_L are fixed, and F and L are variable,

$$\Delta Y = p_F\,\Delta F + p_L\,\Delta L = 0$$

and

$$\frac{\Delta F}{\Delta L} = -\frac{p_L}{p_F}$$

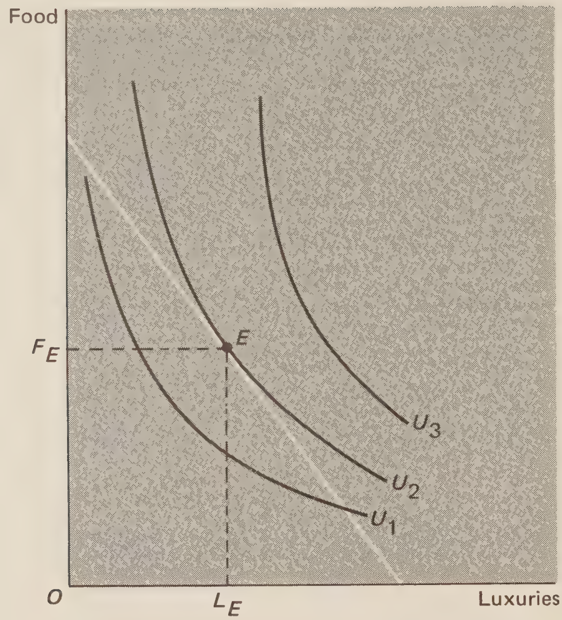


FIGURE 5–13
Indifference Analysis of Consumer Equilibrium

Shifts in the Budget Line

If the consumer's income increases to Y', the budget line would shift to the right, as shown in Figure 5–14. Since the prices of food and luxuries have not changed, the slope of the budget line will not be affected. The increased income allows the consumer to move from U_2 to U_3, a higher indifference curve. The new equilibrium is at G, with spending on luxuries increasing from L_E to L_G. Spending on food increases by a smaller amount, from F_E to F_G. Even though the slope of the budget line remains unchanged, the effect on the quantities of F and L purchased may vary considerably depending on the income elasticity of demand for the respective goods. The quantity purchased of a good may actually decline with income, indicating a negative income elasticity of demand. Such a good is called an **inferior good.**

Changes in the prices of goods can also cause the budget line to shift. Suppose income is Y and the price of food increases. This will cause the slope of the budget line, p_L/p_F, to decrease (in absolute terms). Since p_L and Y have not changed, the amount of luxuries obtainable when no food is purchased is unchanged. Thus the intercept of the budget line with the horizontal (luxuries) axis does not change, and the budget line simply rotates leftward through that point, as shown in Figure 5–15. The price increase causes the individual to move from E on U_2 to D on U_1, a lower indifference curve. Food purchases decline from F_E to F_D, while there is a slight increase in consumption of luxuries, from L_E to L_D.

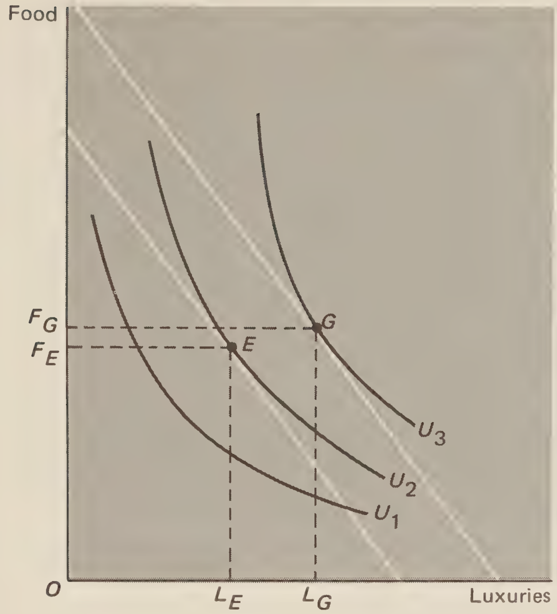


FIGURE 5-14
Effect of Increased Income on Spending

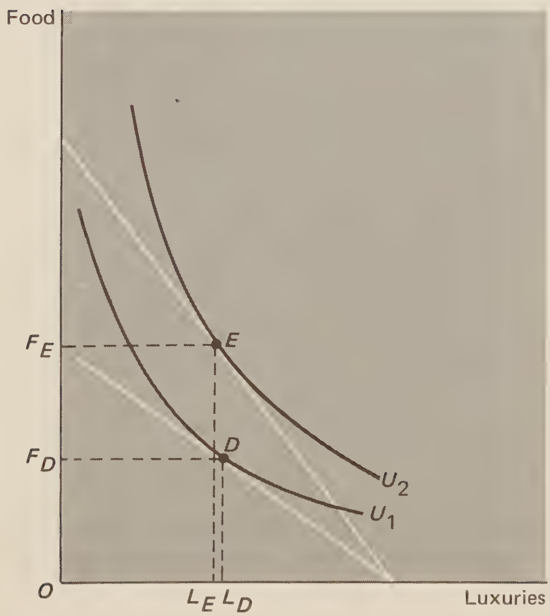


FIGURE 5-15
Effect of an Increase in Food Prices on Spending

Notice that the increase in food prices has caused the consumer to move to U_1, a lower level of welfare and of real income, that is, the price increase has reduced the purchasing power of money income, Y. Is the decline in food purchases attributable to the rise in food prices relative to luxuries or to the fall in real income?

In Figure 5–16 we have drawn a line tangent to U_2 with a slope reflecting the increase in p_F. This is the budget line that would give the consumer his original level of real income or satisfaction after the price change. The consumer would be at equilibrium at C, purchasing F_C of food and L_C of luxuries. The total effect of a change in the price of food on food purchases, $F_E - F_D$, can be broken down into an income effect, $F_C - F_D$, the amount due to a change in real income, and a substitution effect, $F_E - F_C$, the change in food purchases that would take place if the consumer were compensated for his loss in real income. Notice that the substitution effect must always be negative, since the slope of the indifference curve decreases (in absolute value) moving rightward. Thus, as the price line becomes flatter, the point of tangency with the original indifference curve must be to the right of the orignal point, reflecting a lower level of consumption of the good on the vertical axis. The income effect of a price increase is usually negative, too. However, in the case of inferior goods, the income effect will be positive. If the income effect is more important than the substitution effect

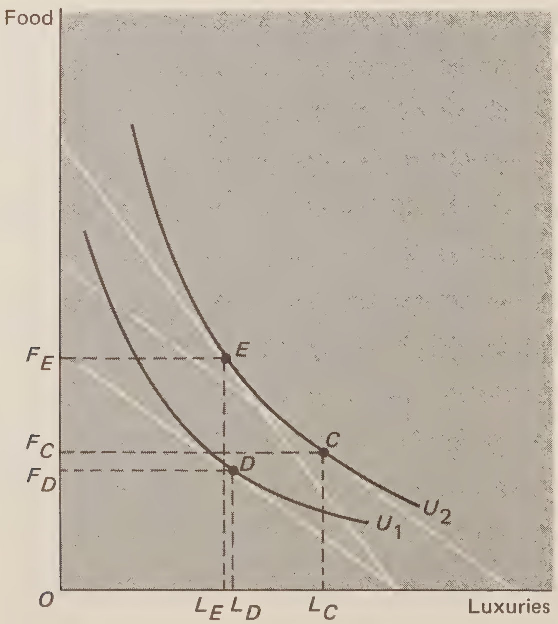


FIGURE 5–16
Income and Substitution Effects of an Increase in Food Prices

and the good is inferior, an increase in its price will cause an increase in the quantity demanded and the good is a Giffen good. This is unlikely to be the case unless the good is a substantial part of the total budget of the consumer.

Designing Cost-of-Living Indices

Many labor union contracts contain **escalator clauses,** which provide automatic wage increases when the consumer price index rises. These are intended to compensate workers for losses in real income associated with increases in the cost of living. If all prices rise proportionately when the consumer price index rises, there will be no problem in determining the correct amount of compensation, since the cost-of-living increase is unambiguous. If all prices rise by 5 per cent, then the cost of living has increased by 5 per cent and a 5 per cent increase in wages would compensate for the loss in real income.

In most cases, however, prices of all goods and services do not rise by the same amount when the consumer price index increases.[9] Those consumers who buy relatively more of the goods with the most rapid price increases will experience a greater decline in real income than the average. Since each worker's indifference map is not known, there is no way to determine the amount of compensation required to regain the former indifference curve. Even if everyone's preferences were known, it would be impractical to provide individualized compensation schemes. Consequently, the amount of compensation required by escalator clauses is designed to provide sufficient purchasing power to obtain the goods and services previously consumed by the average consumer at the new prices. The Bureau of Labor Statistics establishes a typical "market basket" of goods purchased in the base period. If the cost of the market basket of goods rises by 5 per cent, workers receive a 5 per cent increase in money wages.

The amount of compensation is shown in Figure 5–17. The consumer is initially in equilibrium at *E*. If the price of food rises while the price of luxuries remains unchanged, the budget line will shift from *AB* to *AJ*. For the consumer to be able to purchase the original market basket of goods, point *E* must be on the new budget line. Consequently, compensation must move the consumer to line *RS*. Notice, however, that the consumer no longer purchases the goods at *E* (even though they are available to him) but moves to *F* on U_3, a higher indifference curve. Thus money compensation to regain the purchasing power lost as a result of a price change results in an increase in welfare. Therefore, changes in real income (purchasing power) are not an exact measure of welfare change when all prices do not change proportionately.

Although some consumers whose consumption patterns are intensive in those goods and services with the most rapidly rising prices will not be fully com-

[9] The consumer price index is a weighted average of the price changes of a selected set of goods and services, the weights being determined by the proportion of total consumer spending accounted for by spending on the respective goods and services.

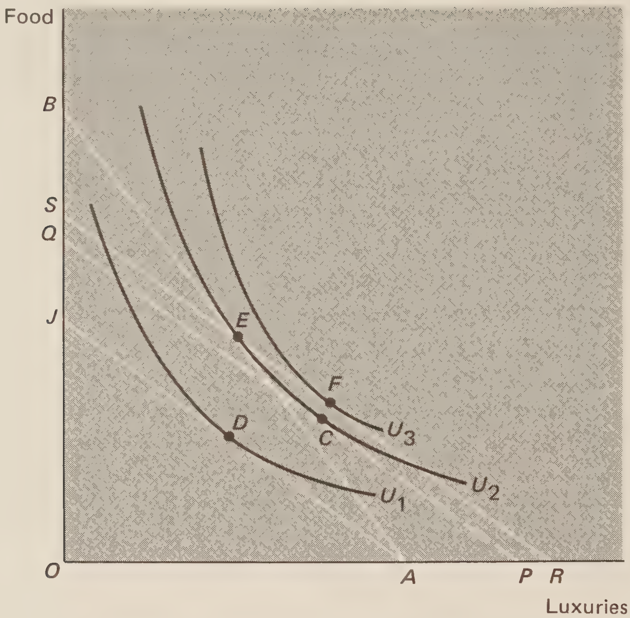


FIGURE 5–17
Impact of Escalator Clauses on Consumer Welfare

pensated for their loss in purchasing power, consumers who purchase the goods and services in the market basket on which the change in purchasing power is reckoned will be better off. Furthermore, if we could compensate each individual for his loss in purchasing power based upon his personal consumption pattern in the base period, all consumers would be better off after compensation than before the price increase.

Derivation of Demand Curves

The classical economists used their psychological model of consumer behavior to derive demand curves for particular goods and services. Demand curves can also be derived from indifference maps. Suppose our consumer must decide how much money to spend on food. Figure 5–18 is the consumer's indifference map between money and food. The slope of the budget line is the price of food, p_F. As p_F decreases, the consumer moves from *A* to *B* to *C*, to successively higher indifference curves, and increases the amount of food purchased. This implies that the demand curve for food is negatively sloped. Furthermore, the amount of money spent on food first increases from ZM_A to ZM_B and then decreases to ZM_C. This is consistent with the declining elasticity of demand characteristic of a typical demand curve.

The line connecting the points of consumer equilibrium associated with

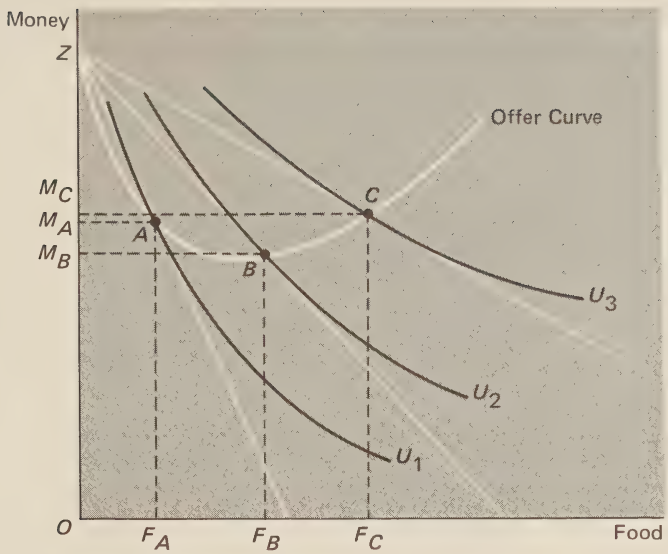


FIGURE 5–18
Derivation of a Demand Curve from an Indifference Map

different prices is called the **offer curve.** A typical offer curve is U-shaped, reflecting the declining elasticity of demand as price falls. When the offer curve is negatively sloped, total money spending on food increases as p_F declines, reflecting demand elasticity greater than 1. When the offer curve is horizontal, total spending does not change, reflecting unitary elasticity of demand. When demand elasticity is less than 1, total spending declines with falling prices and the offer curve has a positive slope. The relation between the offer curve, the demand curve, and the marginal revenue curve is shown in Figure 5–19.

Theory of Exchange

The theory of the household can be used to analyze how product prices are established and to evaluate consumer behavior in terms of the efficiency conditions.

Suppose the United States has a surplus of grain, stockpiled in government warehouses, which it would like to trade to the French, who have a shortage of grain but a similar surplus of wine. The conditions of exchange can be analyzed in the **Edgeworth box diagram** in Figure 5–20.

The left side and bottom are the axes of the wine-grain indifference map for France. $U_1, U_2, \ldots$ are the French wine-grain indifference curves, and O_F is the French offer curve for grain in terms of wine. The right side and top are the axes of the wine-grain indifference map for the United States, turned upside down. $U_1', U_2', \ldots$ are the United States wine-grain indifference curves and O_S is the

United States offer curve for wine in terms of grain. The lengths of the sides represent the actual amounts of grain and wine available for trade.

Trade will occur at some price ratio where both parties are in equilibrium. Since each is in equilibrium only on his offer curve, the trade must occur at *E*,

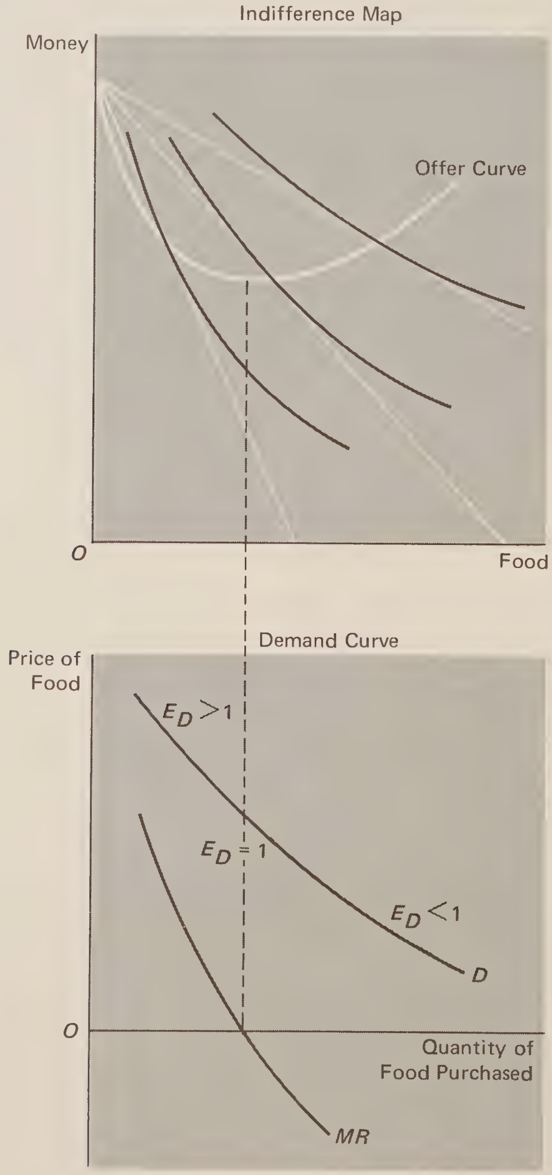


FIGURE 5–19
The Offer Curve, Marginal Revenue, and Elasticity of Demand

where their offer curves intersect. France will trade DW_E units of wine for AG_E units of grain, moving from indifference curve U_1 (at the initial point D) to U_2. The equilibrium price ratio

$$\frac{p_G}{p_W} = \frac{AG_E}{AW_E}$$

is the slope of the indifference curves at E.

Notice that when France is at U_2, the highest attainable indifference curve for the United States is U_3'. Since the trade occurs on U_3' and U_2 at E, the bargain is Pareto optimal, since the United States cannot be made better off without moving to a lower French indifference curve, making France worse off.

Will trade necessarily occur at a Pareto optimal point? Consider all possible Pareto optimal points in the Edgeworth box. These occur where each French indifference curve intersects the highest attainable United States indifference curve, or where the indifference curves are tangent to each other. The locus of these points of tangency is called the **efficiency locus.** Now, we know that each offer curve is the locus of tangency points between the budget line and an indifference curve. Thus, as the price line rotates through point D, it must eventually become tangent to the indifference curves of the traders at some point where the curves themselves are tangent, on the efficiency locus. Since the offer curves of both traders must go through the efficiency locus and since both are U-shaped, they must intersect on the efficiency locus. Because the offer curves intersect there, the efficiency locus is sometimes called the **contract curve.**

As long as the traders view prices as predetermined and are seeking to attain the highest indifference curve, the offer curves reflect the amounts the parties are willing to trade. Thus trade under these conditions will be Pareto optimal.

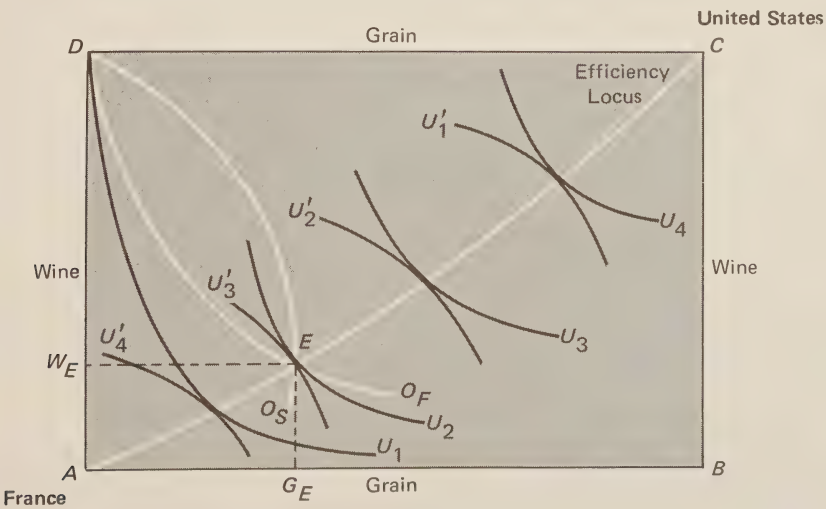


FIGURE 5-20
Exchange of Grain and Wine Between the United States and France

5-3 CLASSICAL AND INDIFFERENCE ANALYSES COMPARED

The classical analysis of household behavior assumed the existence of a quantifiable measure of psychic satisfaction received from goods which could be compared among goods and among individuals. Consumers maximizing utility will purchase goods until the marginal utility of each dollar spent is equal or until

$$\frac{MU_X}{p_X} = \frac{MU_Y}{p_Y} = MU_\$$$

The demand curve for any good is its marginal utility curve (assuming the income effect of a price change is negligible), and the total utility derived from a good is the area under the demand curve. The market demand curve is the horizontal sum of the individual demand curves, and since utility can be compared among individuals, social utility is the area under the market demand curve. Theoretically, the optimal distribution of income can be determined by ascertaining the marginal utility of income, $MU_\$$, for each individual and redistributing income until $MU_\$$ is equal for all individuals.[10]

The indifference analysis drops the concept of cardinal utility altogether and develops a model of household tastes that simply reflects relative preferences among goods. Consumer equilibrium occurs where the budget line is tangent to an indifference curve, which implies that the marginal rate of substitution between Y and X is equal to the ratio of the price of X to the price of Y, that is,

$$\frac{p_X}{p_Y} = MRS_{YX}$$

The marginal rate of substitution is the amount of Y the consumer is willing to trade for X. Moving down an indifference curve, utility must remain unchanged, so that

Utility loss from giving up Y = utility gain from getting more X

or

$$-MU_Y\,\Delta Y = MU_X\,\Delta X$$

Therefore,

$$-\frac{\Delta Y}{\Delta X} = \frac{MU_X}{MU_Y} = MRS_{YX}$$

MRS_{YX} equals the ratio of the marginal utility of X to the marginal utility of Y. If MU_X is 2 and MU_Y is 3, the consumer would be willing to trade $2Y$ for $3X$.

[10] The classical economists recognized the impractiability of determining the optimal income distribution by measuring the marginal utility of income function for each individual in the society. Consequently, they generally assumed that all individuals have the same capacity for psychic gratification from consumption. Since $MU_\$$ is assumed to be a decreasing function of income, this implies that the optimal income distribution is perfect equality, if redistribution has no negative effects on efficiency.

Consequently, the equilibrium condition for the consumer in the indifference approach

$$\frac{p_X}{p_Y} = MRS_{YX}$$

is equivalent to the classical result

$$\frac{p_X}{p_Y} = \frac{MU_X}{MU_Y}$$

The classical analysis assumes diminishing marginal utility for each good. This assumption also implies a diminishing marginal rate of substitution, which is required for the indifference analysis.[11]

Notice that in both cases, as long as all consumers face identical prices, the marginal rate of substitution between commodities is the same for all households. This condition is satisfied regardless of whether goods are purchased from producers at fixed prices or traded between individual households. Thus the condition for Pareto efficiency in consumption is satisfied.

The demand curve for a good in the indifference analysis is derived from the offer curve between the good and money. At all points on the offer curve the price of the good is equal to the marginal rate of substitution between money and the good, that is,

$$p_X = MRS_{\$X}$$

Assuming that the income effect is negligible implies that $MRS_{\$X}$ does not depend on the level of income, so that the indifference curves are vertically parallel, as shown in Figure 5–21. If the price of X rises, the budget line shifts from ZC to ZA and the consumer moves from U_3 to U_2. The demand for X falls from X_C to X_A. Suppose the consumer is compensated for the price increase by an amount ZD, shifting the budget line to YB. Since there is no income effect, the amount purchased at the new, higher price remains X_A and is unaffected by the compensation. Under these conditions, the amount of money compensation required to restore the consumer to U_3, ZD, is a measure of the loss in consumer surplus associated with the price increase.

Interpretation of Market Prices

In the utility model the price a consumer is willing to pay for a good depends on the marginal utility of that good relative to the marginal utility of income, that is,

$$p_X = \frac{MU_X}{MU_\$}$$

---

[11] Diminishing marginal rate of substitution does not imply diminishing marginal utility for each good, but only that the *ratio* of the marginal utility of X to the marginal utility of Y decreases as more X is consumed relative to Y.

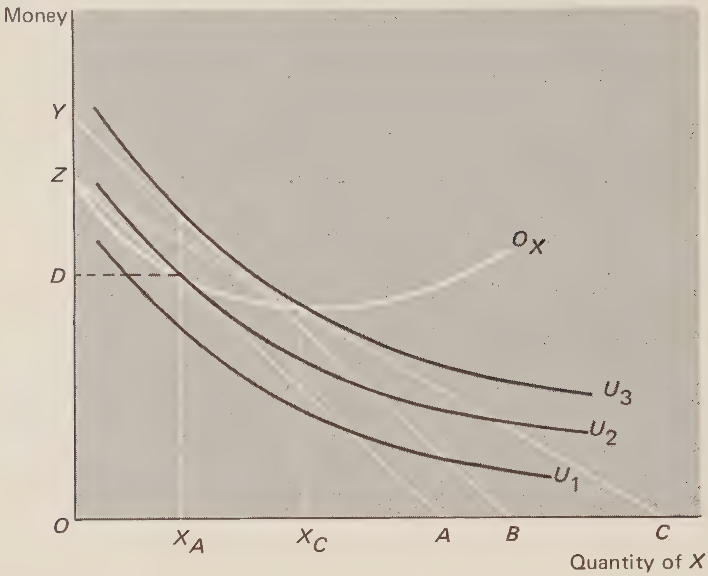


FIGURE 5–21
Demand for a Good with No Income Effect

An individual's valuation of a good, which is implicit in the price he is willing to pay, depends not only on the psychic gratification he receives from consuming the last unit but also on the marginal utility of income. Although two individuals may be willing to pay the same price for a good, the psychic gratification they receive will differ if the marginal utility of income is different for the two. Thus market price reflects the valuation consumers place on a good; however, this may reflect different amounts of utility (psychic gratification) if income distribution is unequal or if the individuals have different capacities for enjoying consumption of goods and services in general.

In the indifference analysis value is always perceived in relative terms. Utility is not an absolute amount of satisfaction but only the amount of satisfaction received from consuming one good instead of another. The marginal rate of substitution reflects the rate at which the consumer would exchange one good for another without gain or sacrifice. Thus, if I am willing to exchange 2 apples for 1 orange so that

$$MRS_{AO} = 2$$

then at the margin I must enjoy the last orange twice as much as the last apple.

If indifference curves are not parallel, preferences depend on income as well as on the relative amounts of goods consumed, since the marginal rate of substitution will vary, even though the relative amounts of the goods consumed stay the same. This is, of course, why it is important to distinguish between income and substitution effects. The income effect measures the change in the con-

sumer's valuation of a good as a result of income change. The substitution effect measures the change in the consumer's valuation of a good as a result of consuming more or less of it relative to some other good (the reference good) than before. Changes in income alone can change market valuation of a good, even though the amount of the good consumed relative to the reference good is unchanged.

The marginal rate of substitution between money and a good can be viewed as the private valuation of a good in the indifference approach, since the consumer adjusts his purchases until

$$p_X = MRS_{\$X}$$

This measures the amount of money a consumer is willing to exchange for the last unit of a good without gain or sacrifice. This interpretation of market price is similar to that in the classical utility analysis. In both cases, "value" in consumption is defined in relation to the consumer's income. Consequently, a knowledge of the incomes of the consumers of a good is important for drawing welfare conclusions from market conditions or making welfare statements about policies that affect market prices.

Welfare Implications

Although the equilibrium conditions and the interpretation of demand curves are analogous for the classical and indifference approaches, the indifference analysis cannot similarly be extended to an evaluation of social welfare. We can analyze the impact of a price change or tax on an *individual's* welfare, but we can say nothing about the *social* gain or loss. Since many price changes affect certain people more than others, this is a serious weakness of the indifference analysis. However, since the assumptions of the classical approach are unrealistic, or at least untestable, the utility analysis cannot be used to determine the social impact of price changes either. At best, we can evaluate whether anyone can be made better off without making someone else worse off, that is, if the change is Pareto optimal.

Suppose the government wishes to raise $100 million to finance a new program of foreign aid. Congress has decided to raise the money by taxes rather than debt issue and seeks the advice of a group of economists concerning the optimal tax policy. The economists are considering two alternatives: a general increase in import duties (a tax on imports) or a proportional income tax. The impact on the balance of payments is not a consideration for this problem, but the tax is to be designed so as to minimize the loss of welfare to consumers.

Indifference curves between imports and money are shown in Figure 5–22. The slope of the budget line *AB* represents the initial price level for imports.[12] The consumer is initially in equilibrium at *F* on U_3.

[12] The price level for imports can be viewed as an index of import prices. We will assume that relative import prices are unaffected by an ad valorem tariff on all imports.

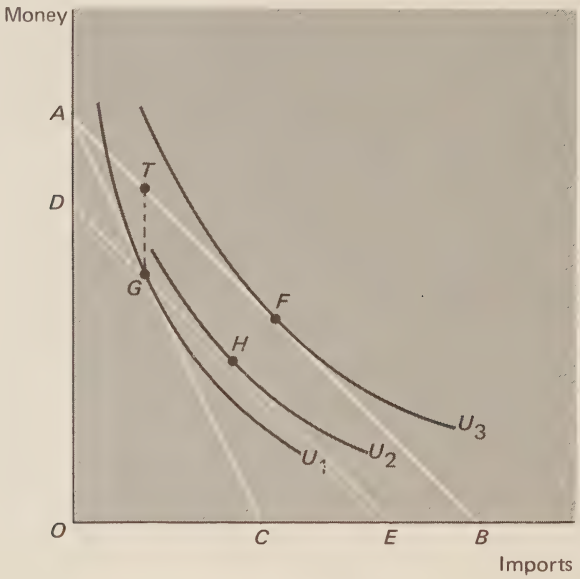


FIGURE 5-22
Comparison of an Import Duty and an Income Tax on Consumer Welfare

If an ad valorem tariff is imposed on imports, the budget line will shift to *AC*. The consumer moves to a new equilibrium at *G* on U_1, a lower indifference curve. Tax receipts from the import duty are measured by *GT*, the difference between the pretariff and posttariff prices at the new equilibrium.

Now suppose the government were to raise an amount *GT* by a general income tax. Such a tax would not affect the price of imports relative to the prices of other goods, so the price line would retain the original slope of *AB*. If the individual pays *GT* in taxes from his income, his budget line would shift to *DE*. His new equilibrium is at *H* on U_2, a higher indifference curve than if the import duty had been imposed.

For any individual, a tax from income will be preferred to a tax on imports that extracts from him an equivalent amount of money. Thus we might conclude that an income tax must be preferred to an import duty when balance-of-payments considerations are ignored.

This is only true, however, when all consumers have similar preference functions. Remember that aggregate tax collections from the two sources must be identical, but this does not necessarily apply to individuals. Those people who are intensive consumers of imports may pay more taxes with an import duty than an income tax. Consequently, they may be worse off with the income tax than with a tariff. If consumption of a good is not uniform (with respect to income) among members of society, then a price change will have distributive consequences that cannot be ignored.

However, a case could be made for the argument that imports in the United States are so diversified that their consumption is sufficiently widespread to make this analysis relevant. If this is true, then to raise $100 million, everyone (or nearly everyone) will be made better off using an income tax rather than an import duty.

5-4 DERIVING A UTILITY FUNCTION FROM AN INDIFFERENCE MAP

The major difference between the classical approach and the indifference analysis is the nature of the model of consumer tastes. The classical analysis assumes the existence of a utility function, while the indifference analysis views consumer preferences in terms of a rank-ordered scale. Two mathematicians, von Neumann and Morgenstern, have shown that a utility function can be derived from ordinal preferences if the consumer can state his preferences in the face of uncertain outcomes.

Suppose you were to bet $10 at 60:40 odds. The mathematical expectation of winning is the probability of the gain less the probability of the loss, or

$$0.6(\$10) - 0.4(\$10) = \$2$$

If you bet $1,000 at the same odds, the expected gain is

$$0.6(\$1{,}000) - 0.4(\$1{,}000) = \$200$$

In both cases, the expected gain is one fifth of the amount wagered. Yet, though you may be willing to take the $10 bet, it is unlikely that you would risk the $1,000 gamble, even though the amount of money you are likely to win is greater. This is because your desire to wager depends not on the actual amount of money you stand to gain or lose but on the utility of the gain or loss. If the marginal utility of income decreases, the more you stand to win, the smaller the gain in utility from the last dollar. On the other hand, as you stand to lose more, the loss in utility rises in greater proportion.

Suppose a person is willing to bet $100 at 60:40 odds but refuses to wager any more money at the same odds. At the margin this implies that the utility of what he expects to gain exactly equals the utility of his expected losses. Since

$$\text{Expected gain} = 0.6(\$100)$$

and

$$\text{Expected loss} = 0.4(\$100)$$

this implies

$$0.6U(\text{gaining } \$100) = 0.4U\ (\text{losing } \$100)$$

or

$$\frac{U(\text{gaining } \$100)}{U(\text{losing } \$100)} = \frac{0.4}{0.6} = \frac{2}{3}$$

Now imagine that to induce this person to bet \$200, he must be given 80:20 odds. In this case,

$$0.8U(\text{gaining } \$200) = 0.2U \ (\text{losing } \$200)$$

or

$$\frac{U(\text{gaining } \$200)}{U(\text{losing } \$200)} = \frac{0.2}{0.8} = \frac{1}{4}$$

Thus the utility of gaining relative to losing an equal amount decreases rapidly the more he stands to lose. This implies that the marginal utility of income is much higher at lower levels of income. The rate of decline can be determined by offering successively more favorable odds to induce the person to wager larger amounts of money. When the person will accept no less than 100:0 odds, the marginal utility of income is infinite.

The von Neumann and Morgenstern approach can be used to measure the relative intensity of wants for goods of a single individual. We could, for instance, offer a person the chance of wining a car at 80:20 odds and determine how much the consumer would be willing to pay for it. In that case, the utility of the car, U_C, could be determined from the equation

$$U(\$) = 0.8U_C$$

where \$ is the amount of money offered. In all cases, utility is relative to the utility derived from other commodities, but the relative *intensity* can be determined even when the level of income has changed. In the indifference analysis, preferences are ranked in such a way that relative intensity can be determined only when the level of income is held constant, that is, for moves *along* an indifference curve. The von Neumann-Morgenstern approach allows us to compare the relative intensity of enjoyment from consuming goods at *different* levels of income.

Although this approach provides information about the intensity of satisfaction a consumer receives from certain goods, nothing can be said about the relative satisfaction experienced by different individuals. It is possible to determine if my marginal utility of income schedule declines more rapidly than yours, but we have no way of knowing if I receive more or less psychic gratification from my income than you do. Thus we are still left without a guide for determining the welfare implications of policy measures that affect income distribution, that is, if they benefit some individuals only at the expense of others.

5–5 THE SUPPLY OF LABOR

Not only do households consume goods but they also supply factor services to firms. Consequently, the theory of the household must include an analysis of the determinants of supply of these services. In this section we concentrate on the

determinants of the supply of labor, but the determinants of the supply of other factors, such as capital, is analogous.

When households supply a factor service to the market, the decision can be viewed as the alternative to consuming some good that provides utility. Work is the alternative to leisure; saving is the alternative to consumption. The theory of labor supply can therefore be viewed as the theory of the demand for leisure.

The problem can be analyzed on the indifference map shown in Figure 5-23. Leisure is on the horizontal axis and income on the vertical. The slope of the budget line *ZA* is the price of leisure in terms of income, or the wage. Equilibrium is at *E,* where the marginal rate of substitution of income for leisure, MRS_{Yl}, is equal to the wage.

Because the alternative to leisure is earning income, the income effect of the work-leisure choice is likely to be quite substantial. At higher levels of income MRS_{Yl} is likely to increase as the marginal utility of income decreases. We would therefore expect households with higher levels of income (or wealth) to supply less labor than poor households for a fixed wage.

Of course, there is a correlation between income and wages. Most poor people are poor because their wages are low and the substitution effect would cause low-wage households to supply less labor and consume more leisure than high-income households. Thus one of the most interesting empirical questions in the analysis of the supply of labor is the relative importance of the income and substitution effects.

Suppose we want to determine the impact of an increase in income taxes on labor force participation. Will it discourage work effort by reducing the wage,

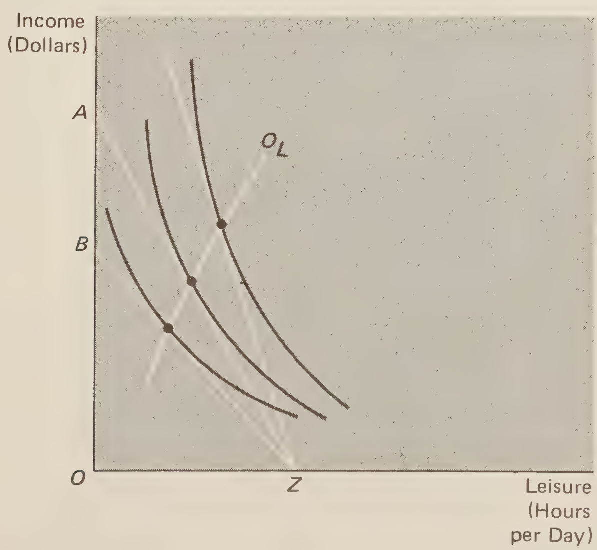


FIGURE 5-23
Indifference Analysis of Labor Supply

or will it cause people to work more hours and impel secondary workers (housewives, teenagers) to enter the labor force to make up for the loss in real income?

In Figure 5–23 the household's preference function shows a dramatic increase in the marginal rate of substitution of income for leisure as income rises. The offer curve is positively sloped over the entire range. As the wage increases, the household increases consumption of leisure and reduces its supply of labor. For this household, the imposition of a tax on income will have the effect of reducing the take-home wage and will cause labor supply to increase.

Such behavior is not necessarily typical. Whether or not the supply of labor will increase or decrease when wages change will depend upon the household's preference function in the relevant range, which will determine the shape of its offer curve. If MRS_{Yl} does not increase as rapidly as income increases, the substitution effect may dominate the income effect and the offer curve will be negatively sloped. In this range an increase in the wage will cause a reduction in consumption of leisure and an increase in labor supply.

Although the nature of consumers' preferences is an empirical question, economic theory would suggest that the income effect is likely to be strongest for low-wage workers, since the marginal utility of income declines most rapidly at low-income levels. When subsistence needs are not gratified, the marginal utility of income is infinite. Once subsistence requirements are met, the MU_Y declines much less rapidly and consequently the MRS_{Yl} increases less rapidly. On the other hand, the substitution effect is likely to be stronger at higher income levels. It is likely to be strongest for those individuals whose consumption of leisure relative to income is low. This implies that the aggregate supply curve for labor, the number of hours worked as a function of the wage, may be "backward-bending," as shown in Figure 5–24. At very low wages a decline in the wage will cause an increase in work effort, because of the income effect, while at high

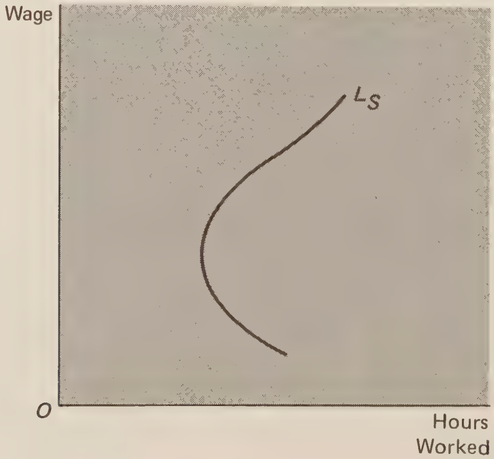


FIGURE 5–24
A Backward-Bending Labor Supply Curve

wages a decline in the wage will cause a reduction in work effort through the substitution effect. Thus, although the actual impact of the tax on work effort must be determined empirically (on the basis of past experience), theoretical considerations can provide useful insights and aid in designing empirical studies.

5-6 REVEALED PREFERENCE

Although the indifference model of consumer tastes is more reasonable than the classical model, many economists are still unhappy with it. To predict a consumer's reaction to changes in price or income, his indifference map must be known. It is unrealistic to assume that a consumer will be able to describe his preferences under all conceivable circumstances, let alone those situations between which he is indifferent.

The **revealed preference** approach does not require any prior knowledge of consumer tastes.[13] If we assume the consumer's tastes do not change over the period during which we study his behavior, we can derive his indifference map.[14] Many of the results derived from the classical and indifference analyses—convex indifference curves, the negative substitution effect of a price change, and so on—can also be demonstrated by revealed preference.

Instead of using a model of tastes to predict consumer behavior, revealed preference uses a model of behavior to describe consumer tastes. Suppose in market period 1 the consumer purchases a set of n commodities $[q^1]$ at prices $[p^1]$. His total expenditure in this market period is

$$E^1 = \sum_{i=1}^{n} p_i^1 q_i^1 = \sum p^1 q^1$$

In market period 2 prices change, and the consumer purchases a different set of m commodities $[q^2]$ at prices $[p^2]$, making a total expenditure of

$$E^2 = \sum_{i=1}^{m} p_i^2 q_i^2 = \sum p^2 q^2$$

Given this market behavior, can we determine whether the consumer prefers $[q^2]$ to $[q^1]$?

The consumer may have purchased $[q^2]$ in period 2 either because he prefers $[q^2]$ to $[q^1]$ or because $[q^2]$ was cheaper than $[q^1]$ in the second period. Suppose

$$\sum p^2 q^2 < \sum p^2 q^1$$

[13] This approach was developed by Paul Samuelson in *Foundations of Economic Analysis* (Cambridge, Mass.: Harvard University Press, 1947). The original treatment uses more difficult mathematics than this formulation.

[14] If tastes change over the period, we cannot talk about the existence of a single indifference map over the period. Constancy of tastes is an implicit assumption of all theories of household behavior.

That is, the value of the goods chosen in period 2 is less than that of the goods chosen in period 1 (valued in period 2 prices). Then we cannot conclude that the individual is better off with $[q^2]$, since he may have decided to purchase $[q^2]$ because it was cheaper than $[q^1]$, not because he preferred it. But if

$$\sum p^2q^2 \geqslant \sum p^2q^1$$

we can say unequivocally that the individual is better off in period 2. That is, he has revealed that he prefers $[q^2]$ to $[q^1]$, since in period 2, $[q^1]$ was as cheap as, or cheaper than, $[q^2]$ and yet he bought $[q^2]$.

Even if the individual experiences no increase in constant dollar income (his expenditure in period 2 prices remains unchanged), so that

$$\sum p^2q^2 = \sum p^2q^1$$

we can still conclude that he is better off in period 2. He has experienced an increase in utility or moved to a higher indifference curve.

The derivation of an indifference map from this model of consumer behavior is shown in Figure 5–25. Suppose the consumer purchases only two commodities,

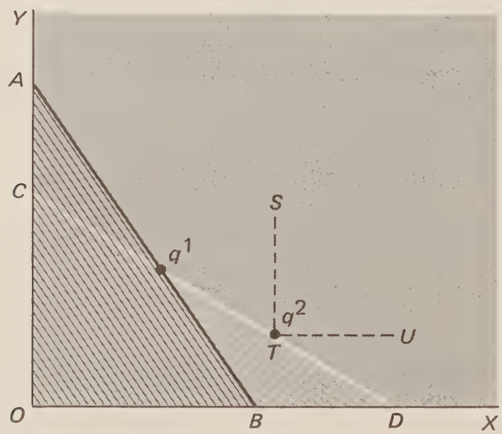


FIGURE 5–25
Derivation of Indifference Map by Revealed Preference

X and Y. The price line in the first market period is AB, and the consumer purchases a combination of X and Y, $[q^1]$. All combinations of X and Y along AB are equally expensive. Now suppose the price line shifts to CD. The consumer will not select a point on CD to the left of $[q^1]$, since these points were cheaper than $[q^1]$ when $[q^1]$ was purchased and they were not selected then.

If $[q^2]$ lies to the right of $[q^1]$ on CD, it must be preferred, since $[q^1]$ is also available. Notice that all points in the shaded region must be inferior to $[q^2]$, since $[q^2]$ is preferred to $[q^1]$ and $[q^1]$ is preferred to all points below AB

(since these were available when $[q^1]$ was purchased) and $[q^2]$ is preferred to all points below *CD*. All points in the region *STU* are preferred to $[q^2]$, since they involve more of both commodities, *X* and *Y*.[15]

If a flatter price line were drawn through $[q^2]$, we could similarly determine if $[q^3]$ were preferred to $[q^2]$. Eventually, by constructing different price lines through each point, we could determine if any point in the space is preferred to, or inferior to, any other point. Thus an indifference map could be constructed.

Relation to Index Number Theory

Samuelson has shown that the theory of revealed preference is similar to index number analysis. When relative prices change over time, it is impossible to separate changes in output from "pure" price increases. Suppose some prices increase by 10 per cent, others by 5 per cent, and money expenditure by 7 per cent. Have the actual amounts of commodities purchased increased or decreased?

Two indices are available. The Laspeyres index measures the change in purchasing power in terms of the base year prices:[16]

$$L = \frac{\sum p^1 q^2}{\sum p^1 q^1}$$

The Paasche index measures the change in purchasing power in terms of the current year prices:

$$P = \frac{\sum p^2 q^2}{\sum p^2 q^1}$$

According to index number theory, real purchases have increased if both *L* and *P* are greater than 1, that is, if

$$L = \frac{\sum p^1 q^2}{\sum p^1 q^1} > 1 \qquad \text{and} \qquad P = \frac{\sum p^2 q^2}{\sum p^2 q^1} > 1$$

This implies

$$\sum p^1 q^2 > \sum p^1 q^1 \qquad \text{and} \qquad \sum p^2 q^2 > \sum p^2 q^1$$

Both of these results are consistent with the theory of revealed preference. The first condition implies that $[q^1]$ was cheaper than $[q^2]$ in period 1. The second condition shows that $[q^2]$ was more expensive than $[q^1]$ in period 2. Since $[q^2]$ was purchased although $[q^1]$ could have been purchased for less money, $[q^2]$ must be preferred to $[q^1]$.

The theory of revealed preference provides some additional insights for in-

---

[15] This indicates that at least near $[q^2]$ the indifference curve must have a negative slope (or it would enter *STU*) and be convex (or it would enter *CD*).

[16] This is the index used to design cost-of-living escalator clauses.

terpreting index numbers in terms of consumer welfare. By the theory of revealed preference, if

$$\sum p^2q^2 = \sum p^2q^1$$

or

$$P = 1$$

the consumer is better off in period 2. This means that an increase in welfare or "real income" has occurred even though the Paasche index of purchasing power has not risen. This result is similar to our earlier finding (from indifference analysis) that using a price index to compensate workers for loss of real income due to inflation will make them better off than before the price change.

Consider the Laspeyres index of price change using base year quantity weights, that is

$$L' = \frac{\sum p^2q^1}{\sum p^1q^1}$$

If that index is used to deflate an index of actual expenditure, E, where

$$E = \frac{\sum p^2q^2}{\sum p^1q^1}$$

then the deflated index

$$\frac{E}{L'} = \frac{\sum p^2q^2}{\sum p^1q^1} \div \frac{\sum p^2q^1}{\sum p^1q^1}$$
$$= \frac{\sum p^2q^1}{\sum p^2q^1} = P$$

is the Paasche index of purchasing power. Thus, if workers are compensated for inflationary price increases as measured by a Laspeyres price index, they will experience an overall welfare gain as consumers.

QUESTIONS FOR STUDY AND REVIEW

1. A consumer's utility function relating the goods A and B is of the form

$$U(q_A, q_B) = q_A + q_B$$

Draw several of the indifference curves. With income and p_B fixed, derive the demand curves for A and B, both as a function of p_A. Suppose that A is rationed, so that the consumer will not consider buying more than r_A at whatever price. Derive the effective demand curves in this situation.

2. A consumer is willing to exchange his automobile for a 20:80 chance of winning a trip around the world. If he loses his trip around the world, a consolation prize is a trip to Florida. Derive the consumer's utility index for his automobile, a trip around the world, and a trip to Florida.
3. Assume that households view the interest rate as a payment for saving. Will an increase in interest rates cause an increase or decrease in household saving? Justify your answer using the indifference analysis.
4. Duesenberry's "relative income hypothesis" states that the utility a household derives from consuming a good depends upon the level of consumption of that good or similar goods by the household's social reference group. How can this model be used to demonstrate that local income taxes should be progressive in relation to income?
5. The committee on culture in a small town has been given the responsibility of operating the community theater. The costs of running the theater depend on the number of people that attend, and these costs can be met either through taxes or admission fees or both. Only members of the community can attend the theater. The community found that if it levies no tax and charges an admission fee of $4.50, each person in the town will attend the theater once per season. If the committee imposes a tax of $2.00 per person and charges an admission fee of $1.50, each person will attend four times; and with a tax of $5.00 and an admission fee of 50 cents, each person will attend seven times. Each of the above proposals covers the cost of operating the theater.[17]
   a) Is the above information sufficient to determine which of the plans should be adopted? If so, what is the best policy for the town; if not, why not?
   b) What are the crucial assumptions in the problem that allow us to base inferences about community welfare on observation of individual market behavior?
6. Tenants in some public housing projects pay about one half the cost of their dwellings. From the point of view of the individual living in a public housing project this arrangement may appear inferior to a comparable income subsidy. Using the indifference analysis, demonstrate why. Can we derive any social welfare implications from this analysis? If so, why? If not, why not?
7. "The existence of income effects limits the usefulness of the theory of the household for analyzing consumer behavior." Carefully explain why this is true.

ADDITIONAL READING

Alchian, Armen. "The Meaning of Utility Measurement," *American Economic Review, XLIII* (March 1953), pp. 26–50.

Barrett, Nancy S. "A Problem in Deducing Welfare Changes from Indices of Real Income" *Journal of Political Economy, LXXVII* (May/June 1969), pp. 434–436.

Baumol, William J. *Economic Theory and Operations Analysis,* 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1972, chapter 10.

[17] This problem is taken from Samuel Bowles and David Kendrick, *Notes and Problems in Microeconomic Theory* (Chicago: Markham Publishing Co., 1970), p. 73.

Friedman, Milton. "The Marshallian Demand Curve," *Journal of Political Economy, LVII* (December 1949), pp. 463–495.

Friedman, Milton, and L. J. Savage. "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy, LVI* (August 1948), pp. 279–304.

Hicks, John R. *Value and Capital,* 2d ed. Oxford: Oxford University Press, 1946, chapters 1–3.

Leibenstein, Harvey. "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," *Quarterly Journal of Economics, LXIV* (May 1950), pp. 183–207.

Robbins, Lionel. "On the Elasticity of Demand for Income in Terms of Effort," *Economica, X* (June 1930), pp. 123–124.

6 The Theory of Production

In the previous chapter we developed a model of household tastes to analyze consumer expenditure patterns. For a given set of preferences, the pattern of consumption depends on the prices at which goods and services are offered as well as on the individual's income. To find the terms on which goods and services will be offered to consumers by producers, we need a theory of production. Such a theory focuses on the supply side of commodity markets. Since production theory also deals with the process of factor valuation and remuneration, it provides the basis for a theory of individual income distribution. By providing a framework for analyzing supply prices and household incomes, the theory of production is a necessary complement to the theory of consumption and to the analysis of the satisfaction of human wants.

It was noted that there are alternative institutional arrangements a society may evolve for satisfying human wants with limited resources. Economists have traditionally assumed that the models of household preferences developed in chapter 5 are universally applicable. Consumers in any society will wish to maximize the utility or satisfaction they receive from their incomes. While preference functions as well as income distribution may differ among societies, the principles of consumer choice, given preference patterns and income, are the same. The major difference among economic systems is in the institutional arrangements under which production takes place.

The theory of production can be divided into two parts. First, we can develop a model of the technology of production that is independent of the institutional arrangements for production. Second, we will consider the way in which differences in producers' objectives affect production decisions.

6–1 TECHNOLOGY AND OPTIMIZATION IN PRODUCTION

An economic system is concerned with the best use of scarce resources. Consequently, any producer, regardless of the economic

system in which he operates, will always prefer the technique that gives the greatest output for a given input. Equivalently, he will prefer a technique that produces a given output with the least input cost. Using this postulate, we can develop a number of conditions for optimization in production.

The model of production used in economic theory assumes that the firm has available to it a known **technology** represented by a **production function.**[1] A production function relates quantities of input (or factors) to an output.

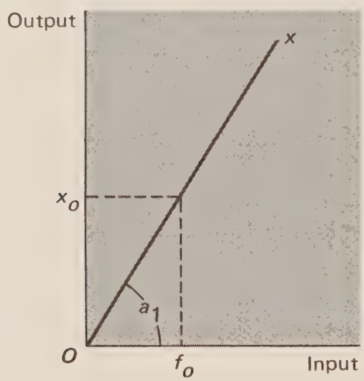


FIGURE 6–1
Input-Output Production Function

The simplest production function is the input-output relation shown in Figure 6–1. Algebraically, the function can be written

$$X = aF$$

where X is output and F is input. If input is F_0, output is X_0.

Alternatively, the production function can be viewed as an input demand function. Suppose you want to produce X_0 of output. Then F_0 would represent the input requirement. Algebraically, input demand is a function of output, that is,

$$F = \frac{X}{a}$$

If we know the cost at which the firm obtains the input, then the input demand function can be viewed as a cost function, that is, the cost of production as a function of the level of output. Suppose F is man-hours of labor and w is the hourly wage. Then production cost as a function of output is

$$\frac{wF}{X} = \frac{w}{a}$$

[1] The actual technology available to firms may vary among economic systems just as consumer preference functions vary. While this is important for analyzing *particular* production and consumption decisions in particular situations, the actual empirical specification of production and preference functions is not necessary to the theoretical analysis.

The cost function associated with the input-output production function is shown in Figure 6–2. Cost per unit, or **average cost,** is constant for the input-output relation, that is,

$$\frac{wF}{X} = \frac{w}{a_1}$$

The average cost function is a horizontal line.

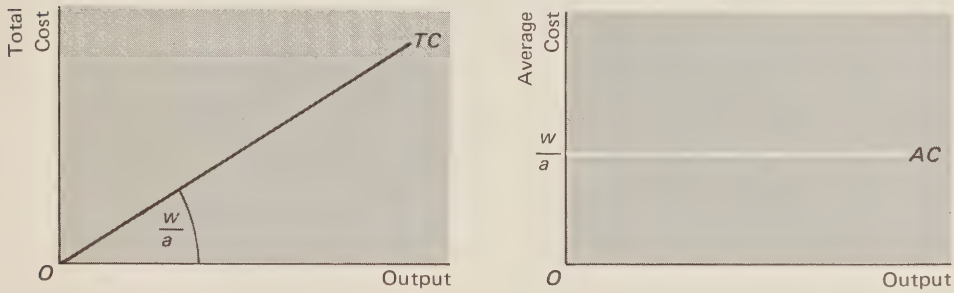


FIGURE 6–2
Input-Output Cost Function

This simple linear production function is very restricted. There is only one factor that produces the output. The relation between the factor and its product is exactly proportional. This property is called **constant returns to scale,** or **linear homogeneity.** It implies that the ratio between the factor and its product is invariant to the scale of operation.

Extension to Two Inputs

In most cases, there are several inputs required to produce output. Suppose there are two inputs: L, labor, and K, capital, which are combined to produce an output, X. 2 L and 3 K are required to produce 1 X. The process can be repeated at higher levels so that 4 L and 6 K will produce 2 X; 6 L and 9 K will produce 3 X; and so on. Such a process is a technique for producing X and is called an **activity.**

In Figure 6–3 the activity is shown graphically. It is a straight line from the origin with a slope of 2/3, the ratio at which L and K are used to produce X. The amounts of X produced are read as indices along A_1. For instance, combining 4 L and 6 K produces 2 X. The activity exhibits constant returns to scale, since we assumed that the process can be repeated at any level, so the output indices are equally spaced along A_1.

There are usually a number of ways in which resources can be combined to produce the output; technologically, this implies that there are several activities available to the firm. The firm must decide which activity (or combination of activities) should be used. Consider the production function shown in Table

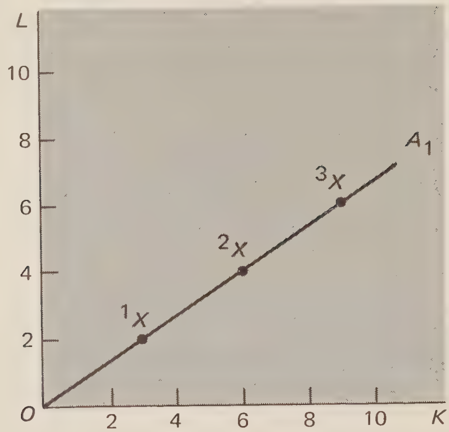


FIGURE 6–3
Activity Analysis of Production

6–1. It can be viewed as a set of activities, as shown in Figure 6–4. The corresponding activity matrix is Table 6–2. Three different processes are available to the firm, each using K and L in different proportions. Activity 1, A_1, which uses a large amount of K relative to the other activities, is called K-intensive, while A_3 is L-intensive.

TABLE 6–1
Production Function for a Hypothetical Firm

Output	Input L	Input K
1	2	3
1	2.5	2
1	5	1
2	4	6
2	5	4
2	10	2
3	6	9
3	7.5	6
3	15	3

Suppose the firm wishes to produce 2 units of X. The locus of all combinations of K and L that can be used to produce a stipulated amount of X is called an **isoquant.**[2] The 2-unit isoquant is shown in Figure 6–4. Points on the activities that represent 2 units of output are on the isoquant as well as lines connecting

[2] Notice the similarity between the isoquant (constant output locus) of production theory and the indifference curve (constant utility locus) of the theory of the household.

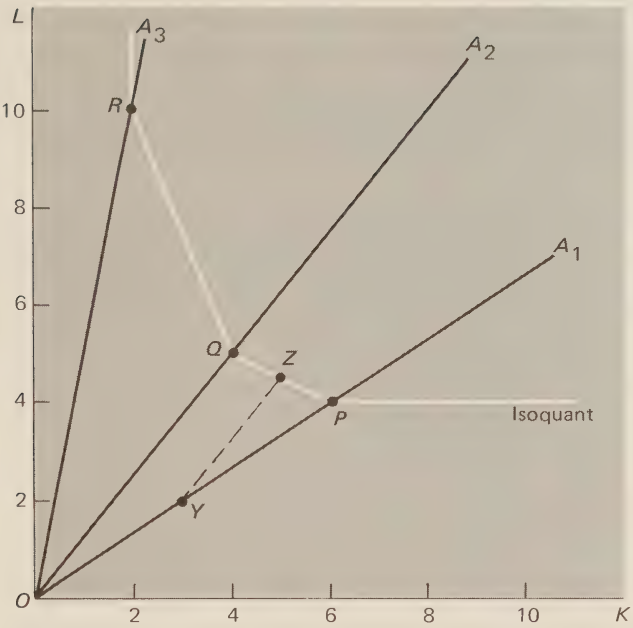


FIGURE 6-4
The Production Function as a Set of Activities

those points. Firms can produce output by a combination of activities, and these combinations result in points on lines connecting the activities. Suppose, for instance, the firm produces 1 unit of X using A_1 and 1 unit using A_2. OY is one half of the distance to the 2-unit isoquant on A_1. Since the remaining X is produced by A_2, YZ must be parallel to A_2 and one half of OQ. Since two sides of YZP are proportional to those of OQP, the third side must also be proportional and the triangles must be similar. This requires that ZP coincide with QP. We could similarly show that any point on the line connecting Q and P represents a combination of A_1 and A_2 and must be on the 2-unit isoquant.

It can also be shown that the isoquant must be convex to the origin. Say that we combine A_1 and A_3, the most L-intensive and K-intensive activities. Points on a line connecting P and R would also be on the 2-unit isoquant. However, K-L combinations on this line would require more K and L than points on PQR. Since

TABLE 6-2
Activity Matrix for a Hypothetical Firm

A_1	A_2	A_3	Resources
2	2.5	5	L
3	2	1	K

points on the line *PR* are always available, an input combination above *PR* would never be selected, so an efficient isoquant would never be concave.[3]

A complete production function contains an isoquant for each level of output, as shown in Figure 6–5. Because we have assumed that the available activities are invarient to the scale of operation, the isoquants will be equally spaced and

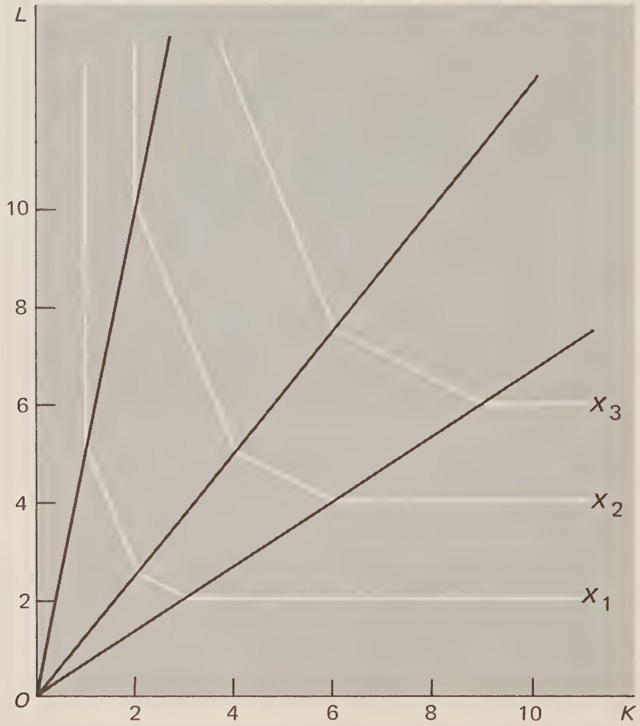


FIGURE 6–5
The Production Function as a Set of Isoquants

parallel with respect to the origin. They will be equally spaced because the amounts of *K* and *L* required to increase production along any activity (or using any fixed combination of activities) will increase in exactly the same proportion as the output. They are parallel because the slope of the lines connecting the activities depends upon the location of the equal output points on the activities. Since the proportional increase in input is the same for all activities, the relative position of the equal output points remains the same, so the slope of the isoquants is invariant to the scale of operation. If the isoquants are equally spaced and parallel with respect to the origin, the production function is said to be linear homogeneous and it exhibits constant returns to scale.

[3] If the points lie on a straight line (for instance, if A_2 were not available), the isoquant is said to be **weakly convex.**

The Neoclassical Production Function

If we could combine L and K in any way with continuously varying outputs, the isoquants of such a production function would be as shown in Figure 6–6. We call such a production function neoclassical because of its obvious similarity to the indifference map of the neoclassical theory of the household.

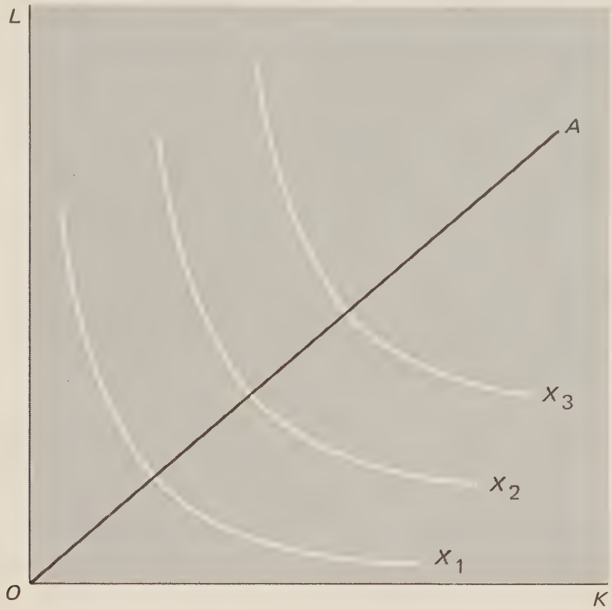


FIGURE 6–6
A Neoclassical Production Function

Notice that the slope of these neoclassical isoquants is constantly decreasing (in absolute value). This implies that each activity is truly independent in the sense that it cannot be viewed as a combination of other processes.[4] However, the neoclassical approach to production theory makes no specific assumptions about the spacing or the slopes of the isoquants. Suppose a set of neoclassical isoquants is equally spaced and parallel with respect to the origin. This implies that along a ray from the origin (line A in Figure 6–6), when the factors are combined in the same ratio, increases in inputs produce proportionate increases in output. Thus the production function displays the property of linear homogeneity.

The property of linear homogeneity can be stated mathematically in the following way. Suppose output is a function of L and K, that is,

$$X = f(L, K)$$

[4] Since the slope decreases at each point, a neoclassical isoquant is everywhere strictly convex. The isoquants in Figure 6–5 are weakly convex in some regions.

The function is linear homogeneous if a proportional increase in each input causes output to increase proportionately, that is,

$$\lambda X = \lambda f(L, K) = f(\lambda L, \lambda K)$$

where λ is the factor of proportional increase.

A mathematical form of a linear homogeneous production function that is commonly used in empirical research is the Cobb-Douglas function:

$$X = AL^{\alpha}K^{1-\alpha}$$

where α is a constant reflecting individual factor productivity[5] and A is a constant measuring the overall efficiency of the technology.

Suppose both L and K are increased by a constant proportion, λ. Performing the appropriate multiplications and reducing the exponents on λ,

$$A(\lambda L)^{\alpha}(\lambda K)^{1-\alpha} = A\lambda^{\alpha}\ \lambda^{1-\alpha}L^{\alpha}K^{1-\alpha} = \lambda AL^{\alpha}K^{1-\alpha} = \lambda X$$

we find that the effect is to multiply output X by the same proportion, λ. Thus the Cobb-Douglas production function is linear homogeneous.

Nonconstant Returns to Scale

Suppose the isoquants are not equally spaced, as in the production function shown in Figure 6–7. As inputs are increased along activity A the increment in output decreases. Such a production function exhibits decreasing returns to scale or homogeneity of a degree less than 1. If the isoquants move closer together as output expands, the production function is said to exhibit increasing returns to scale or homogeneity of a degree greater than 1.[6]

For many purposes, the assumption that a typical production function is linear homogeneous (homogeneous of degree 1) is not unrealistic. It simply means that the efficiency of a process is independent of the scale of operation and that any process can be used at any scale of operation. Clearly, however, a problem arises when a process requires a factor of production that is available only in large units. Chamberlin argued that some processes require specialized equipment that is so costly that it only becomes useful when large outputs are required.[7] A research professor in a small economics department may rely solely upon students operating desk calculators to perform his statistical analysis. In a large university he may have a high-speed electronic computer available. From his point of view, student statistical work is more costly in terms of time and probability of error than using a computer, but it would be costly to acquire an IBM 370

---

[5] The constant α is technically the elasticity of output with respect to labor input, or the ratio of per cent change in output to per cent change in labor input.

[6] A function $X = f(L, K)$ is homogeneous of degree n if

$$\lambda^n X = \lambda^n f(L, K) = f(\lambda L, \lambda K)$$

[7] See Edward H. Chamberlin, *The Theory of Monopolistic Competition*, 8th ed. (Cambridge, Mass.: Harvard University Press, 1962), appendix B.

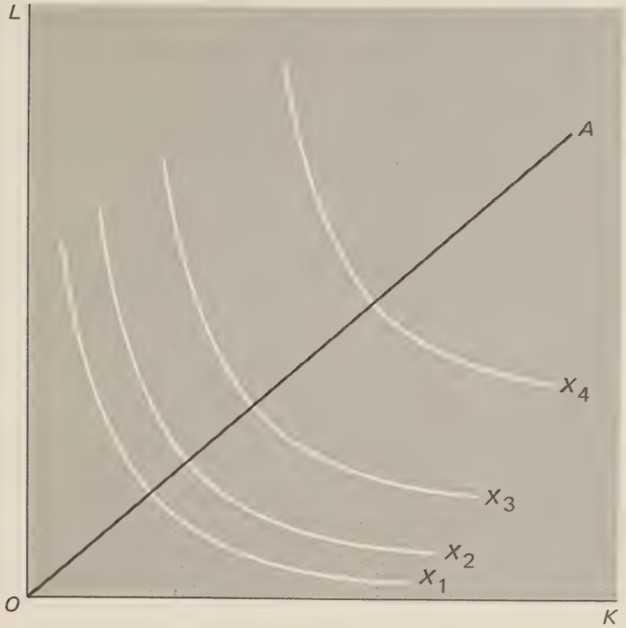


FIGURE 6–7
Neoclassical Production Function Exhibiting Decreasing Returns to Scale

for a single research project. Because the cost of the IBM 370 can be spread over more research projects, a large university will be willing to use such a process. Clearly, however, the computerized process is only cheaper for large scales of operation.

Chamberlin also points to increased specialization and division of labor at higher scales of operation. At a large university it pays to train an economics student in computer programming and employ him to provide technical assistance to research projects. A smaller university may have a single student assistant trained in programming to assist all users of the computer. Such a student would be less helpful to the individual researcher, since the student would lack competence in the subject area of the research.

Empirical research has shown that for many large firms which have presumably attained the optimal long-run scale of operation and which are large enough to take advantage of the most efficient production techniques, production functions are approximately linear homogeneous, implying constant returns to scale.[8] In empirical models of the technology of the American economy as a

[8] One of the first such studies was Martin Bronfenbrenner and Paul Douglas, "Cross Section Studies in the Cobb-Douglas Function," *Journal of Political Economy, XLVII* (December 1939), pp. 761–783. Other studies however, have found both increasing and decreasing returns to scale in different industries. See A. A. Walters, "Production and Cost Functions," *Econometrica, XXXI* (January 1963), pp. 1–66.

whole, where the aggregate production function is viewed as the sum of the production functions of individual firms, the assumption of linear homogeneity is generally made.[9]

Optimal Production Decisions

Regardless of whether the firm's production function consists of a small set of potential activities or the possibilities for factor substitution are very great, the manager must decide the best mode of production. The optimal factor mix will depend upon the availability of the factors in relation to the production function.

Suppose there are only three independent activities available to the firm for producing its output X. The isoquant map is shown in Figure 6–5, and the production function and activity matrix are shown in Tables 6–1 and 6–2. If the firm has available to it 11 units of L and 13 units of K, how much output can be produced and which activities will be used to produce it? In Figure 6–8 the isoquant map is reproduced. The shaded area $OKEL$ represents all the feasible production possibilities given the factor availabilities. Point E represents the highest level of output consistent with the availability of resources, since it lies on the highest isoquant in the feasible region. Optimal production would involve using a combination of A_1 and A_2. Point E can be reached by moving along A_1 to isoquant 3 and then moving along FE, parallel to A_2, to isoquant 5. This implies that process A_1 is being run at a scale of 3 and A_2 at a scale of 2 to produce 5 X.

In most cases, the firm is able to purchase additional resources in factor markets, or in the case of a firm in a socialist state, it can obtain resources at a fixed accounting price from the central planning board. Using the technique of **linear programming,** we can determine whether additional L or K should be purchased. Suppose the firm sells its output at a fixed price per unit, p_X. We know that the resource cost of producing a unit of X using A_1 is 2 L and 3 K, that is,

$$C_1 = 2L + 3K$$

while the resource cost per unit using A_2 is 2.5 L and 2 K, that is,

$$C_2 = 2.5L + 2K$$

If the firm can purchase a unit of L for p_L and K for p_K, then the cost functions become

$$C_1 = 2p_L + 3p_K \qquad \text{and} \qquad C_2 = 2.5p_L + 2p_K$$

We are interested in determining the real value of our inputs L and K for producing X. In other words, we want to know the opportunity cost of L and K in terms of X. If we purchase one more unit of L, how much additional X will we obtain? The opportunity cost of L and K in terms of X is called their **shadow price.**

[9] For example, see Robert M. Solow, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics, XXXIX* (August 1957), pp. 312–320.

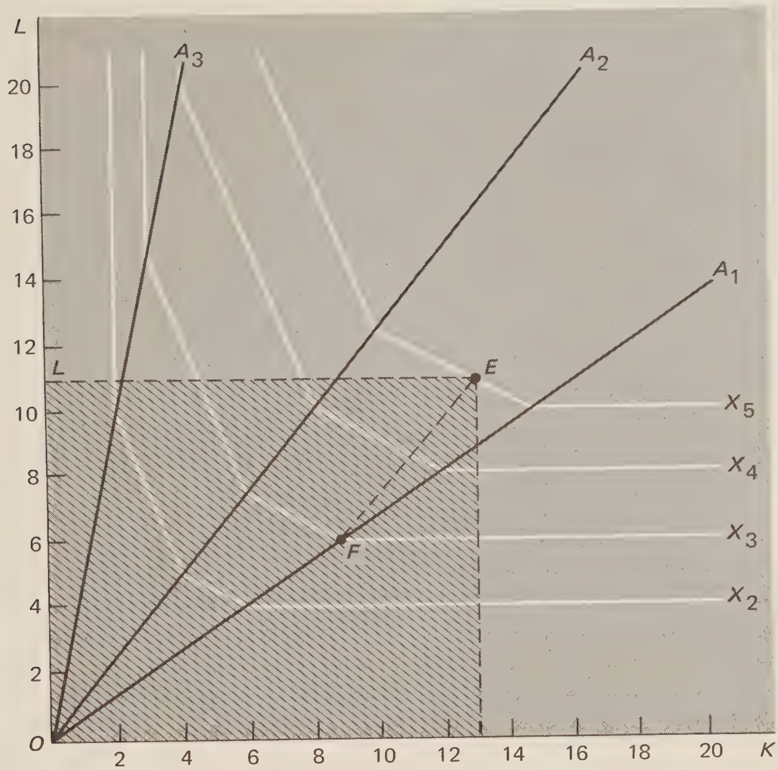


FIGURE 6-8
Activity Analysis of Optimal Production Techniques

Shadow prices are obtained by imputing to the resources the value of the product. For A_1, the value of output p_X can be allocated to the inputs in terms of their relative proportions in the process, that is,

$$p_X = 2\pi_L + 3\pi_K$$

where π_L is the shadow price of L and π_K is the shadow price of K. For A_2,

$$p_X = 2.5\pi_L + 2\pi_K$$

Since both activities are used, the shadow prices can be obtained by solving the equations simultaneously. Thus,

$$\pi_L = \frac{2}{7} p_X \qquad \text{and} \qquad \pi_K = \frac{1}{7} p_X$$

If p_X is \$1, the imputed value of L is 28 cents per unit and the imputed value of K is 14 cents per unit. Since the shadow price represents the amount the factor actually contributes to production, this means that if the market value of either input is less than the shadow price, the producer will find it advantageous to

purchase more of the input. If the market price is greater than the shadow price, the producer will find it advantageous to release some of his factors.[10]

It is interesting to note that we have obtained prices or values for resources from purely technological information. This remarkable property of linear programming is called **duality.** Any time we must optimize the use of scarce resources, shadow prices for those resources can be obtained from the optimal solution. Prices are a necessary by-product of any economizing decision and therefore must exist (even if only implicitly) in any economic system regardless of its institutional arrangements.[11]

Use of Marginal Analysis in the Neoclassical Model

If the production function has the neoclassical property of strictly convex isoquants (implying an infinite number of independent activities), the marginal analysis can be applied to an analysis of optimal production techniques. Consider the neoclassical isoquant map in Figure 6–9. Suppose the firm can acquire resources (from the market or the board) at fixed prices, p_L and p_K. The amount of L and K obtainable for a fixed cost is given by line AB, derived from the equation

$$C = p_L L + p_K K$$

The line has a slope equal to the negative of p_K/p_L, that is,

$$-\frac{\Delta L}{\Delta K} = \frac{p_K}{p_L}$$

In other words, the price ratio is the rate at which K can be exchanged for L without affecting total cost.[12] Point E on isoquant 2 represents the highest level of output obtainable for the stipulated level of cost. Alternatively, AB is the lowest cost for which 2 X can be produced.

The optimal production technique occurs at a point of tangency between the cost line and an isoquant. The slope of the isoquant is the rate at which L can be exchanged for K without affecting the level of output. This is the **marginal rate of**

[10] In a socialist state factors will be released to the board, since the accounting values will exceed the actual contributions to production. In a market system resources will be attracted to the factor market, since the firm will not wish to pay the factors as much as the prevailing market rates.

[11] For a more technical discussion of linear programming techniques and duality see William J. Baumol, *Economic Theory and Operations Analysis,* 3d ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1972), chapters 5 and 6.

[12] The slope can be derived in the following way. If L and K are to be exchanged without affecting cost, and assuming p_L and p_K remain fixed,

$$\Delta C = p_L\,\Delta L + p_K\,\Delta K = 0$$

Therefore,

$$-\frac{\Delta L}{\Delta K} = \frac{p_K}{p_L}$$

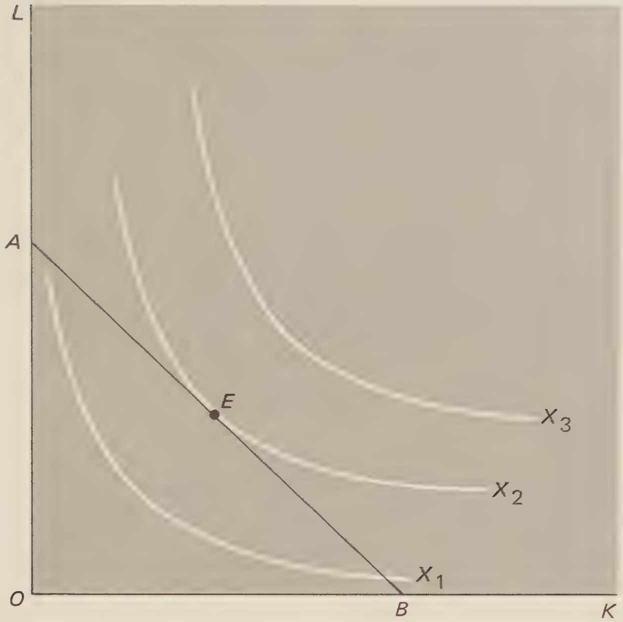


FIGURE 6-9
Marginal Analysis of Production

factor substitution, MRS_{LK}. Thus the condition for efficiency in production is

$$\frac{p_K}{p_L} = MRS_{LK}$$

We can look at the marginal rate of factor substitution in another way. Consider the impact of an increase in L on output, with K held constant. Since K and L can be combined in any proportion, this will normally produce an increase in output. The ratio $\Delta X/\Delta L$ is called the marginal product of L, MP_L.[13] Moving up an isoquant, output must remain unchanged, so that

$$\begin{matrix}\text{reduction in output} \\ \text{from reducing } K\end{matrix} = \begin{matrix}\text{increase in output} \\ \text{from increasing } L\end{matrix}$$

or

$$-MP_K \, \Delta K = MP_L \, \Delta L$$

Therefore,

$$-\frac{\Delta L}{\Delta K} = \frac{MP_K}{MP_L} = MRS_{LK}$$

[13] In activity analysis the marginal factor products are generally zero, since there are only a finite number of independent ways to combine K and L. Consider the impact of a unit increase in K on the example in the text. Unless L also increases, there can be no increase in output.

Say that MP_L is 3 and MP_K is 2. Then if L is substituted for K, there is a gain of 3 X per unit of L and a loss of 2 X per unit of K. To hold output constant 2 L would replace 3 K, since there would be a gain of 6 X associated with the change in L and a loss of 6 X associated with the change in K. Therefore,

$$MRS_{LK} = \frac{MP_K}{MP_L} = \frac{2}{3}$$

Another way of stating the efficiency condition in production is that the ratio of factor prices must be equal to the ratio of their marginal products, that is,

$$\frac{MP_K}{MP_L} = MRS_{LK} = \frac{p_K}{p_L}$$

Imputation of Value

Suppose the firm can sell its output at a fixed price, p_X. The opportunity cost of a resource is the amount of revenue forgone if a unit is removed from production. The amount of output forgone is the marginal product of the factor. The revenue forgone is the marginal product multiplied by the product price, sometimes called the marginal value product. Thus the imputed value of each factor will be its marginal product multiplied by p_X, that is,

$$\pi_L = p_X \times MP_L$$

where π_L is the shadow price of L. In the neoclassical model resources will always be paid their shadow prices or, equivalently, their marginal value products. Since

$$\frac{\pi_K}{\pi_L} = \frac{p_X(MP_K)}{p_X(MP_L)} = \frac{MP_K}{MP_L} = \frac{p_K}{p_L}$$

it follows when market prices are equal to shadow prices resource utilization is optimal.

Activity Analysis and the Marginal Approach

There has been considerable debate recently concerning the merits of the neoclassical approach with respect to production theory. Critics argue that the assumption of activity analysis, that is, that there are a limited number of independent technological processes available to a firm, is more realistic than the neoclassical model.

The neoclassical model provides a theory of production that is analogous to the theory of the household. Once the indifference analysis has been mastered, marginal analysis of production follows easily. Isoquants are similar to indifference curves; the cost line is equivalent to the budget line. The conditions for equilibrium, both graphically and algebraically, are analogous. It would be a simple matter, for instance, to analyze the impact of changes in factor prices and

in total cost outlays on production decisions once one has mastered the effect of changes in product prices and total income on consumption decisions.

If the neoclassical analysis can be viewed as a reasonable approximation to reality, its results should be consistent with those of a more realistic model such as the finite-activity production function. To the extent that these results are consistent, the neoclassical approach is useful for economists who are accustomed to the techniques of marginal analysis.

There are a number of properties held by the models in common. Duality, the ability to obtain shadow prices on the basis of technological considerations, is an important common result. Furthermore, the shadow prices of linear programming can be interpreted as marginal products in the sense that they measure the opportunity cost of the resource in terms of the output. The notion that optimal resource utilization, or equivalently the demand for the factor, depends upon shadow prices (marginal products) is also common to both approaches.

We shall see, however, that the neoclassical assumption that factors can be substituted for each other in any way leads to the conclusion that factor price adjustment is an appropriate mechanism for eliminating unemployment of a factor. If technology is limited to a finite set of activities, this will not be the case. Consequently, activity analysis will provide different conclusions from the neoclassical approach.

In sum, many of the results of linear programming, or activity analysis, are equivalent to the marginal analysis, so that the neoclassical model is a useful approximation to that approach. On the other hand, as we evaluate the neoclassical model we will see that it sometimes produces results that are no longer valid when the assumptions of the model are violated. Consequently, it is important to recognize the purposes for which marginal analysis is useful for a model of production theory and when it is not.

Homogeneity and Optimal Factor Proportions

Suppose a firm is contemplating an expansion in the scale of its operations. Factors of production, L and K, can be obtained from the market or the CPB at fixed prices p_L and p_K. The technology is represented by the isoquant map in Figure 6–10. We have assumed that the production function is linear homogeneous, so the isoquants are equally spaced and parallel with respect to the origin.

As the firm expands to a larger scale, total cost increases. Since p_L and p_K remain fixed, the slope of the cost line is unaffected by the expansion, shifting from AB to CD to EF. Points of equilibrium are found where the shifting cost line is tangent to successively higher isoquants. The locus of these points is called the **expansion path.** Now, since the isoquants of a homogeneous production function are parallel with respect to the origin, the locus of their tangency with a line that shifts parallel to the origin must be along a straight line. Thus

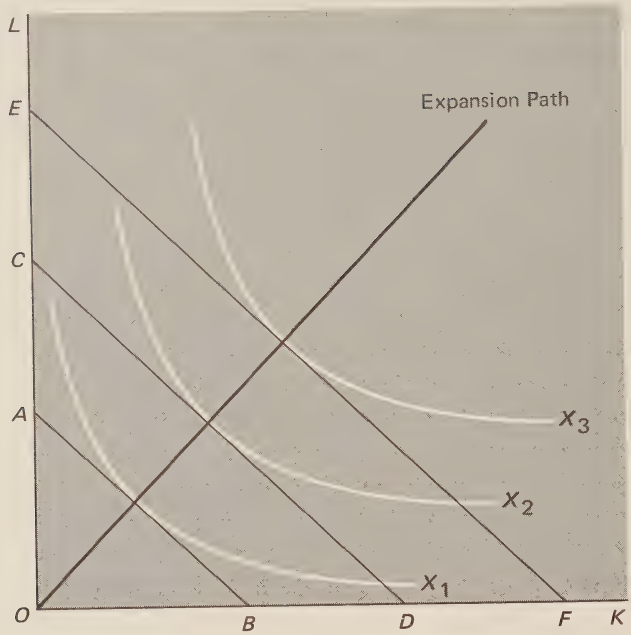


FIGURE 6–10
Expansion Path for a Homogeneous Production Function

homogeneous production functions are characterized by linear expansion paths.[14]

Along a linear expansion path the ratio of L to K is unaffected by the scale of operations. This implies that the optimal factor proportions, given constant factor prices, are the same regardless of the size of the enterprise. This is true for any homogeneous production function, even if returns to scale are not constant.

In Figure 6–11 we have a nonhomogeneous production function with a nonlinear expansion path. Notice that as production increases from X_1 to X_2, the optimal factor proportions become more capital-intensive. At still higher levels of output production becomes more labor-intensive.

If the production function is linear homogeneous, and the isoquants are strictly convex, the law of diminishing returns must hold when one factor is varied and the other remains fixed. Suppose we hold the amount of K constant at K_0 and vary L along AK_0 as shown in Figure 6–12. Because we have assumed constant returns the isoquants are equally spaced. Since the slope of the isoquants are equal along E, the slope of B must be steeper than at D. Consequently, the distance BC must be greater than CD. This implies that it takes more L to move from X_2 to X_3 than it takes to move from X_1 to X_2. Since the change in output is

[14] Homogeneous production functions have linear expansion paths regardless of the degree of homogeneity. The degree of homogeneity reflects returns to scale, while the homogeneity property indicates the proportions in which factors are used at different scales of production.

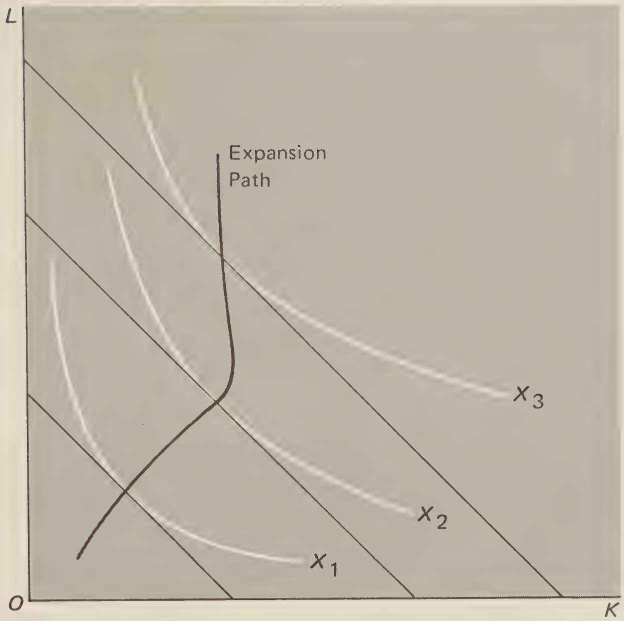


FIGURE 6-11
Expansion Path for a Nonhomogeneous Production Function

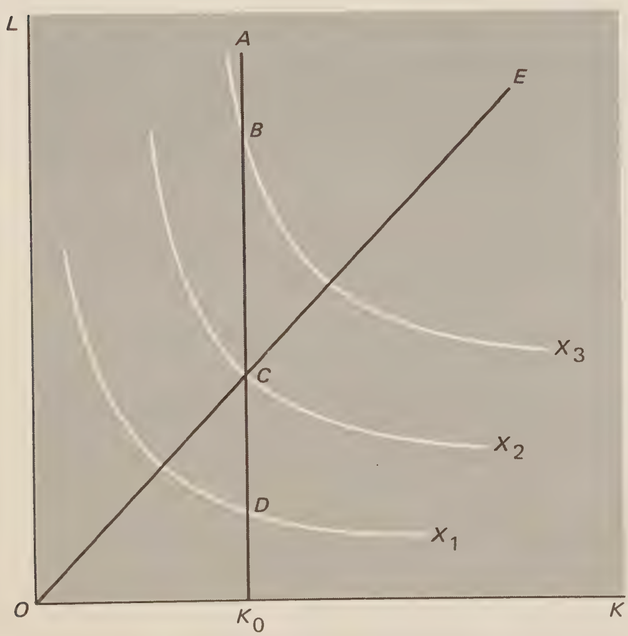


FIGURE 6-12
Diminishing Returns and Linear Homogeneity

the same for both moves, the marginal product of L is declining at higher output levels.[15]

6–2 COST CURVES AND PRODUCTION FUNCTIONS

Once the firm has determined the optimal factor proportions for each scale of production, a cost function can be derived. **Total cost** is the cost of production for each level of output. Since the firm is presumed to have selected the optimal factor proportions, this can be viewed as the least cost function. **Average cost** is cost per unit of output, and **marginal cost** is the cost of producing an additional unit. Both average and marginal cost can be derived from the total cost function.

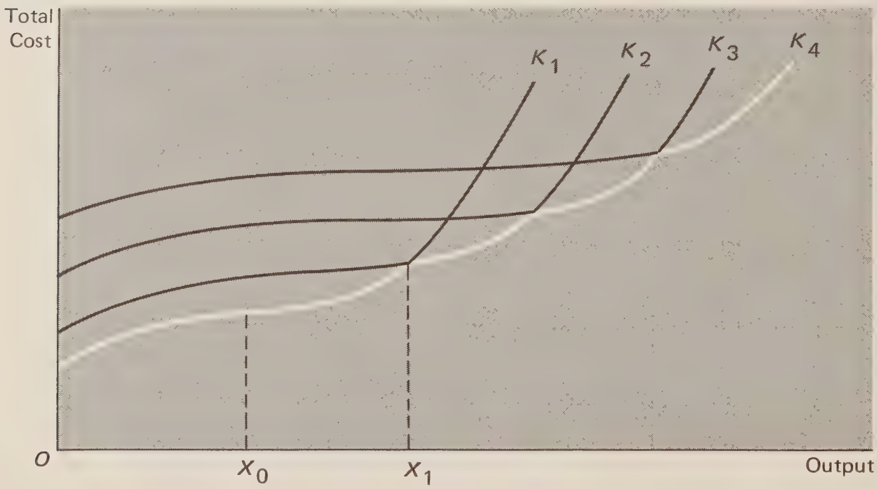


FIGURE 6–13
Short-Run and Long-Run Total Cost Function (Linear Homogeneous Case)

We can distinguish between short-run and long-run costs. In the short run some factors of production, capital equipment, plant, management, and so on, are assumed to be fixed, so that the firm cannot increase output along its expansion path. In the long run, however, all factors are variable, so that the cost function can be derived from the expansion path. In any event, with factor prices given, the cost function can be viewed as a mirror image of the production function. In Figure 6–13 we have shown short-run total cost functions (*STC*) for a firm for four different plant sizes, expressed as different values of K. Assuming that factor prices are fixed, each *STC* function exhibits increasing and then de-

[15] This result can be demonstrated for any homogeneous production function displaying nonincreasing returns to scale. If returns to scale are increasing, then X_3 will be closer to X_2 than X_2 is to X_1, so that the scale factor may more than compensate for the increasing marginal rate of substitution of L for K.

creasing returns to the variable factor, L. That is, costs rise as more labor is hired, but at first at a decreasing rate. Since p_L is fixed, the marginal product of L is increasing. At X_0 on $K=1$ the marginal product of L begins to decline, since costs are increasing more than proportionately to output. At output X_1 it becomes desirable to switch to a larger plant $K=2$. The envelope of the STC curves shown by a heavy line represents the lowest cost of producing X when plant size is a variable. This represents the long-run total cost curve, LTC. If K were available in very small units, LTC would be a straight line from the origin. With fixed factor prices and a linear homogeneous production function, long-run total cost should increase in proportion to total output.

Marginal, Average, and Total Cost

Marginal and average cost can be derived from the total cost function. In Figure 6–14 we have shown the relation between the short-run cost functions.

In the short run it is useful to distinguish between total cost, TC, and total variable cost, TVC. TVC is total cost less fixed cost. The TVC function is parallel to the TC function, with its origin at zero. When output is zero, variable costs are zero by definition. The distance between TC and TVC on the vertical axis represents fixed costs.[16]

Marginal cost, the change in cost per unit of output, or

$$MC = \frac{\Delta TC}{\Delta X} = \frac{\Delta TVC}{\Delta X}$$

is the slope of the total cost function. Since TC and TVC are vertically parallel, their slopes are the same for all values of X, and MC is the slope of either function. As output rises to X_1 the slope of TC is decreasing, so MC is decreasing. Beyond X_1, MC increases as the slope of TC increases. Minimum MC occurs at X_1, the inflection point of TC and TVC.

Average cost is the ratio of total cost to output, that is,

$$AC = \frac{TC}{X}$$

Average variable cost is the ratio of total variable cost to output, that is,

$$AVC = \frac{TVC}{X}$$

Because marginal cost declines and then rises with output, the AC and AVC functions must also be U-shaped. Minimum AC is found at X_3 where a line from the origin is tangent to TC. The slope of a line from the origin is TC/X, or AC, and the tangency point X_3 occurs where the line with the lowest slope (AC) intersects

[16] In the long run all costs are variable, so that TC is equivalent to TVC.

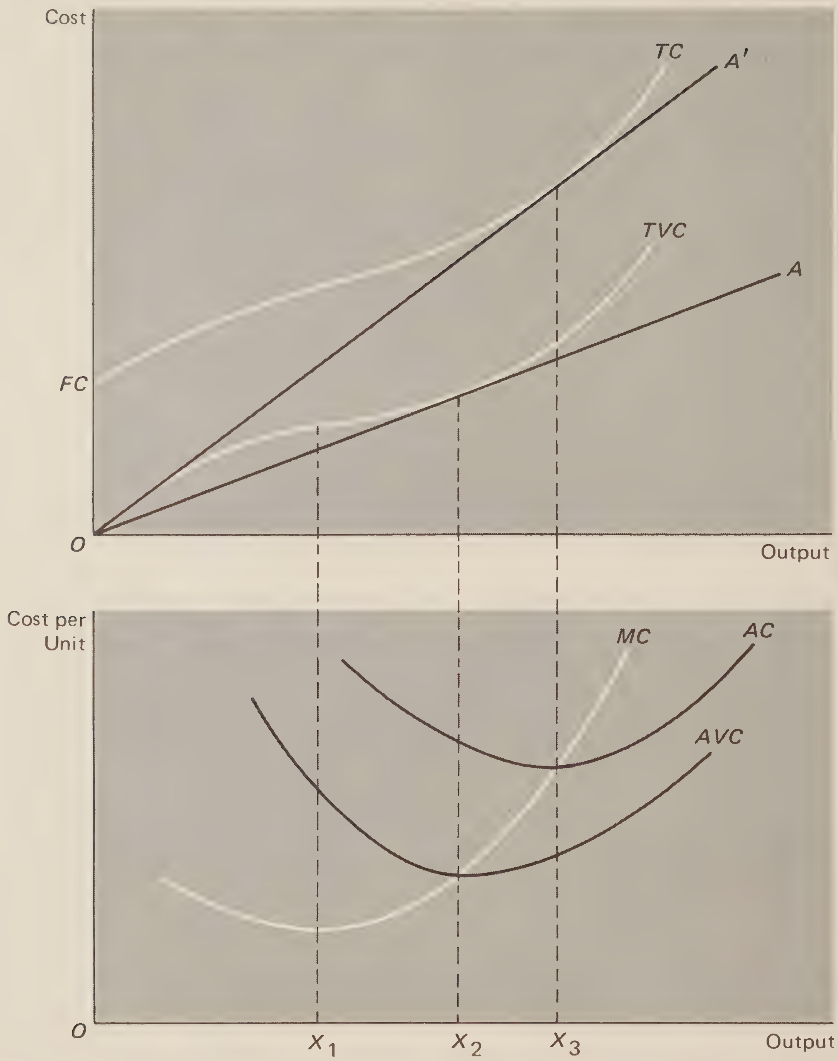


FIGURE 6–14
Derivation of Marginal and Average Cost from the Total Cost Function

TC. Equivalently, minimum *AVC* is found at X_2 where a line from the origin is tangent to *TVC*.

Notice that *MC* intersects *AC* and *AVC* at their minimum points. This is because *MC* is the tangent to *TC* and *TVC* and so must be identical to *AC* at X_3 and to *AVC* at X_2. This result can also be obtained in the following way. As long as the addition to total cost (or *TVC*) is less than average cost (or *AVC*), average cost (or *AVC*) must be decreasing. If *MC* exceeds *AC* (or *AVC*), then *AC* (or *AVC*) must be increasing.

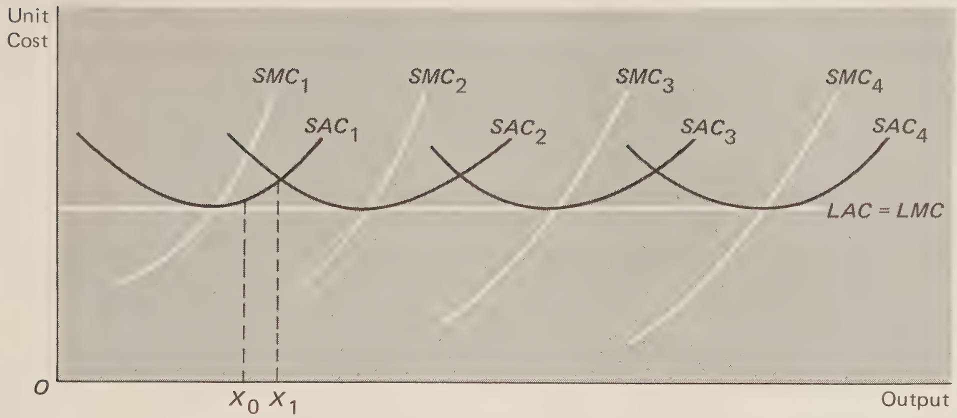


FIGURE 6–15
Short-Run and Long-Run Average and Marginal Cost Functions (Linear Homogeneous Case)

The relation between the long-run marginal, average, and total cost curves is similar. In the case of constant returns, long-run average cost is constant. This implies long-run average cost (*LAC*) must be equal to long-run marginal cost (*LMC*), which is also constant.

The relation between long-run and short-run cost curves is shown in Figures 6–15 and 6–16. Even if costs are constant in the long run, short-run production costs (per unit) may be rising as output increases. If factors are available only in large units, the firm may produce in the rising portion of its *SAC* curve indefinitely, depending upon demand conditions. In Figure 6–15 the firm would not switch to a larger plant until demand conditions warranted an expansion of output beyond X_1. At X_0 the firm would be producing under conditions of rising unit cost, but

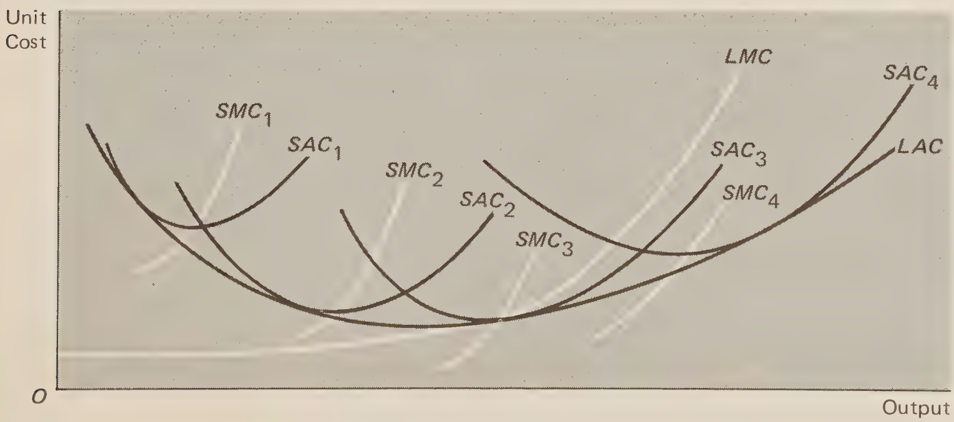


FIGURE 6–16
Short-Run and Long-Run Average and Marginal Cost Functions (Nonconstant Returns)

it would not be profitable to switch to a larger plant. Thus, unless all factors are available in very small units, unit costs of production may be nonconstant with respect to output, even though the underlying production function is linear homogeneous.

Nonconstant Long-Run Costs

In Figure 6–16 the long-run average cost function is U-shaped. Nonconstant long-run average costs may be due either to changes in factor prices as output expands or to an underlying production function that displays increasing and then decreasing returns to scale.[17]

Factor prices may decline at first as the firm expands and is able to employ more effective purchasing or recruiting arrangements. But as the firm continues to expand, factor prices may rise as factor supplies are depleted. Less efficient workers, poorer material resources, or inferior equipment may be put into use.

Some economists argue that the typical long-run *AC* curve is likely to be U-shaped because of increasing and then decreasing returns to scale in the production function. Increasing returns are due to the ability to utilize specialized factors of production that are only available in large units. But at very large scales of operation organizational and managerial problems will produce inefficiency.

Empirical studies indicate that for many industries *LAC* does decrease over a substantial range.[18] But there is no evidence of rising *LAC* at large scales of operation. Thus the typical *LAC* curve, at least for manufacturing industries, tends to be L-shaped rather than U-shaped. This does not necessarily imply that organizational difficulties do not plague management of large corporations. Indeed, the fields of management science, operations research, systems analysis, and so on, have grown *pari passu* with the growth of industry. In addition, most large corporations decentralize divisional management to such an extent that divisions in the same company produce competing products. This is particularly true for automobiles, cigarettes, soap, and packaged food products.

The establishment of competitive divisions within large firms is understandable in the light of the efficiency of the competitive model for resource allocation. It is interesting to note that many firms use systems of accounting prices for internal allocation of resources, much as does the board of the competitive socialist model.

6–3 TRANSFORMATION OF COMMODITIES

So far we have focused attention on the production of a single commodity. In a general model of resource allocation we must be concerned with production

---

[17] It is important to recognize, however, that if *LAC* is U-shaped, it would be due to scale factors. The U shape of *SAC* is attributable to diminishing returns to the variable factor when others are held fixed.

[18] See Walters, "Production and Cost Functions."

alternatives. If the government is to purchase ten additional F-111 jet fighters, what will be the cost to the private sector? If resources are fully employed to begin with, they must be released from other endeavors to the aircraft industry. The amount of other commodities that must be foregone to produce an additional F-111 is called the **marginal rate of transformation.**

We can examine the question of commodity transformation through the eyes of a producer of a multiple product line. Fresh Home Products Corporation produces both bath soap and toothpaste. If production is efficient, the firm will want

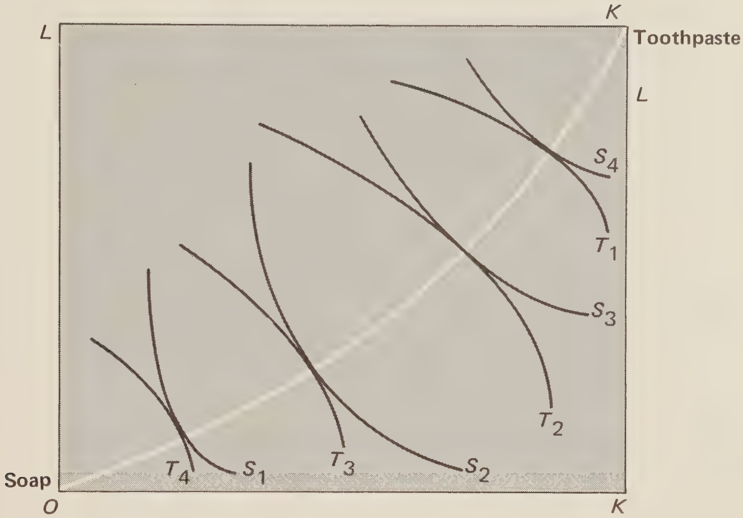


FIGURE 6–17
Efficient Production Alternatives

to maximize its output of soap production, given the level of toothpaste production, or alternatively, to maximize toothpaste production, given the level of soap output.

Assume for a moment that the optimal scale of operations has been determined, so that the amounts of L and K available to the firm are fixed. The problem can be analyzed by means of the familiar Edgeworth box diagram in Figure 6–17. The isoquant map for soap is represented on the southwest corner of the box, while the toothpaste isoquants are superimposed (upside down) on the northeast corner. The lengths of the sides of the box represent the amounts of L and K available to the firm.

Suppose the firm wishes to maximize soap production when toothpaste output is T_3. From the Edgeworth box, this occurs at S_2 where T_3 is tangent to a soap isoquant. Notice that T_3 can be viewed as the maximum output of toothpaste when soap production is at S_2. For any level of toothpaste production, the highest level of soap output is found at a point of tangency between a soap isoquant and

a toothpaste isoquant. A curve through these tangency points is called the **efficiency locus.** Regardless of the desired output mix, production should take place somewhere on the efficiency locus.

Since efficient production occurs at points of tangency between isoquants for different products, their slopes are equal there. Recall that the slope of an isoquant is the marginal rate of factor substitution, or

$$-\frac{\Delta L}{\Delta K} = MRS_{LK} = \frac{MP_K}{MP_L}$$

Thus, along the efficiency locus, the marginal rate of factor substitution must be the same for both commodities:

$$MRS_{LK}^S = MRS_{LK}^T$$

where superscripts S and T refer to soap and toothpaste, or

$$\frac{MP_K^T}{MP_L^T} = \frac{MP_K^S}{MP_L^S}$$

This implies that the ratio of the marginal factor products must be equal for both goods, that is,

$$\frac{MP_K^T}{MP_K^S} = \frac{MP_L^T}{MP_L^S}$$

The ratio of factor products is the marginal rate of transformation. Suppose one unit of L is transferred from toothpaste to soap. Then the change in T relative to S will be

$$-\frac{\Delta T}{\Delta S} = \frac{\Delta T/\Delta L}{\Delta S/\Delta L} = \frac{MP_L^T}{MP_L^S}$$

Since

$$\frac{MP_K^T}{MP_K^S} = \frac{MP_L^T}{MP_L^S}$$

the same would be true if K were transferred from one product to another.

A **transformation curve** shows the maximum amount of one commodity that can be produced, given some fixed level of production of the other commodity. A transformation curve can be derived from the Edgeworth box as shown in Figure 6–18. The maximum amount of T that can be produced for any level of S can be read from the isoquant indices along the efficiency locus. For instance, when soap production is S_2, the maximum level of toothpaste output is T_3.

The absolute value of the slope of the transformation curve is the marginal rate of transformation, MRT_{TS}, where

$$MRT_{TS} = -\frac{\Delta T}{\Delta S} = \frac{MP_K^T}{MP_K^S} = \frac{MP_L^T}{MP_L^S}$$

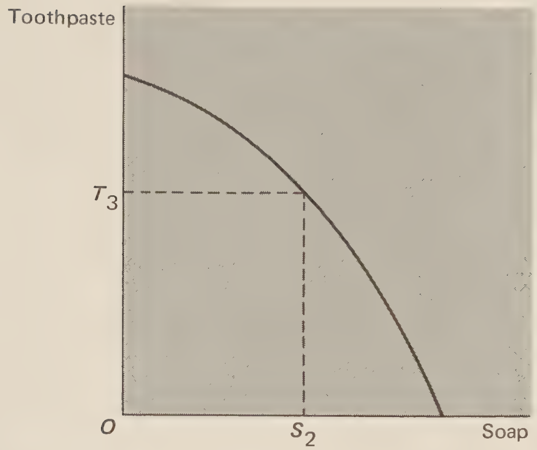


FIGURE 6–18
A Transformation Curve

The marginal rate of transformation can be viewed as the opportunity cost of soap in terms of toothpaste. If both production functions display nonincreasing returns, the transformation curve will be concave to the origin. This means that the opportunity cost of either good in terms of the other rises the more of the good is produced.[19]

The transformation curve is defined for a given technology and fixed factor supply. Suppose a decline in wages causes the firm to increase the amount of labor employed. The impact on the transformation curve is shown in Figure 6–19. The increase in L causes an increase in the length of the L axes of the Edgeworth box, shown by the dashed lines in Figure 6–19. If toothpaste production is maintained at T_3, soap production can rise from S_2 to S_3. At T_4 soap production can rise from S_1 to S_2. Equivalently, if soap production is S_2, toothpaste output can rise from T_3 to T_4.

The effect of an increase in one factor of production on the transformation curve will depend on the relative factor intensities of the goods. Since soap is more K-intensive than toothpaste, soap uses less L per unit of K than toothpaste. Consequently, the shift along the S axis will be less than along the T axis, as shown in Figure 6–19.[20] For any level of toothpaste production, T_0, the toothpaste cost of soap rises. In other words, toothpaste, the L-intensive commodity, becomes cheaper in terms of soap, the K-intensive commodity, as L increases relative to K.

[19] In the production function shown, soap is K-intensive relative to toothpaste. As more soap is produced less K is available for toothpaste production and toothpaste is produced by more and more L-intensive activities. Less toothpaste can be produced per unit of L than before, so that the amount of toothpaste foregone (the toothpaste cost of soap) must rise.

[20] Since the absolute amount of the shift depends on the units in which S and T are measured, we must refer to relative shifts. The shift will occur in such a way as to flatten the transformation curve, regardless of the units of measurement.

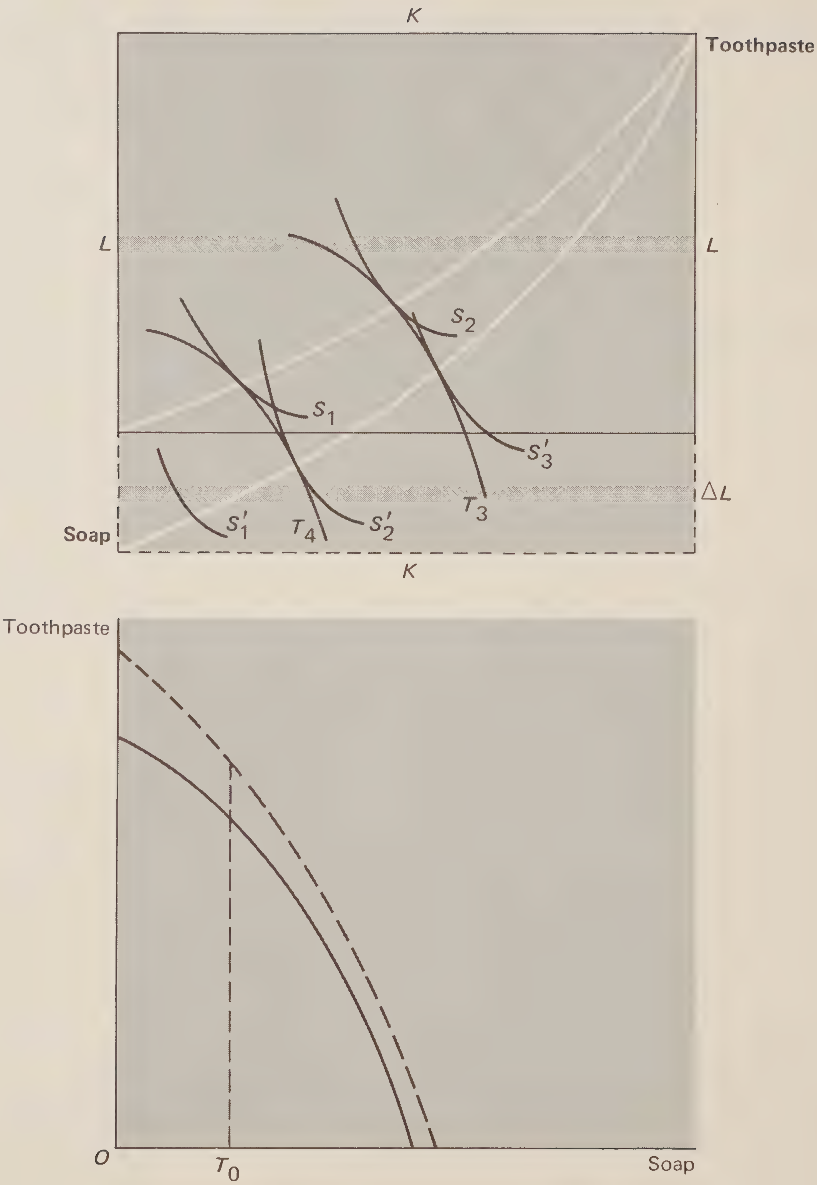


FIGURE 6–19
Impact of a Change in Factor Supply on the Transformation Curve

This property of the transformation curve has important implications for the theory of specialization and exchange. It explains why factor proportions play such an important role in determining the international patterns of production. Countries with abundant labor and scarce capital should specialize in labor-

intensive commodities, because the opportunity cost is lower than for countries with a scarcity of labor. As factor proportions change in the course of development, the pattern of production should change as well.[21] For example, Japan in the 1950s was principally a producer of labor-intensive goods. Growth in capital stock occurred rapidly with little growth in the labor force. By the 1960s Japan began exporting steel, automobiles, electronics equipment, and other capital-intensive items. As growth occurred, the opportunity cost of these items relative to labor-intensive goods declined dramatically.

The Factor Proportions Problem: An Application

An important distinction between the neoclassical approach and the assumption of finite production activities is in the implications for resource utilization. Given factor availabilities, all factors will be employed in production along the efficiency locus when the isoquants take the neoclassical form, that is, when they are strictly convex. Thus full employment of all resources is assured at all points on the transformation curve.[22]

But suppose technology is limited to a finite number of independent activities. Consider the case of a developing country, Pacifica, which has just received technological assistance from a team of United States advisers. Prior to the arrival of the advisers the principal economic activity of Pacifica was subsistence agriculture, a highly labor-intensive operation. Manufactured goods were imported from abroad. The advisory team now proposes the introduction of modern agricultural techniques as well as the establishment of some manufacturing industry. The technology of the industrial sector is to be imported from the United States and consequently is capital-intensive.

The impact on the technological options of Pacifica is shown in Figure 6–20. The dimensions of the Edgeworth box represent the factor endowments. A_L is the original labor-intensive agricultural process; A_K is the modern agricultural process; and M is the single process for producing manufactured goods. If a tariff is imposed on manufactured goods that compete with domestic products, this will have the effect of stimulating demand for domestically produced manufactured goods in Pacifica. As manufacturing output increases from zero, the efficiency locus is the line OF and agricultural production becomes more and more labor-intensive, that is, less agricultural output will be produced by the new modern technology. This is because more and more of Pacifica's scarce

[21] It is often difficult to effect changes in the structure of production in response to changes in factor proportions, because of vested interests in the old industries. As opportunity costs rise in the old industries, managers complain that factor prices are too high, blaming unions and high taxes for rising costs. Clearly, however, the problem lies in their rising transformation cost relative to other goods, not in the absolute level of factor prices or taxes, which presumably are the same for all industries.

[22] This assumes, of course, that factor prices are set at a level that will clear the factor markets. Unemployment may result from barriers to factor price reductions that cause supply to exceed demand. The neoclassical analysis implies that all unemployment can be eliminated by appropriate reductions in factor prices.

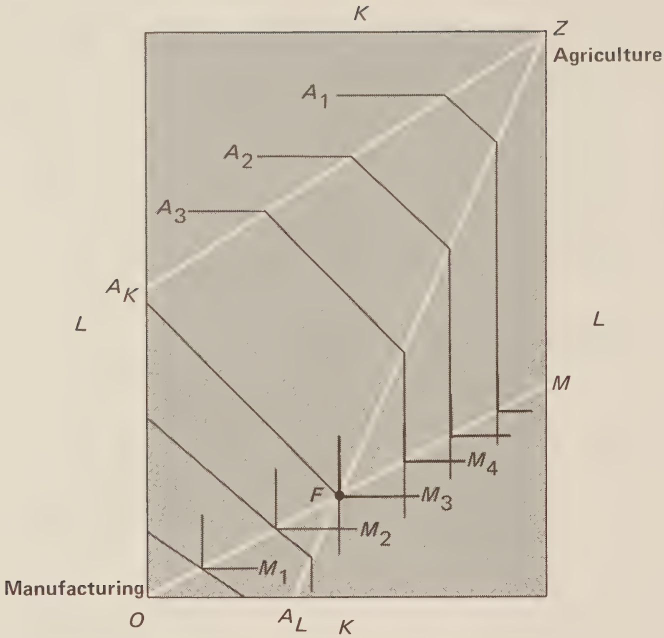


FIGURE 6–20
The Factor Proportions Problem

capital is used in manufacturing and less is available for agriculture. If manufacturing output is to exceed M_3, the efficiency locus is the area FZM. Production takes place using A_L and M, with the vertical distance between A_L and M representing unused labor, or unemployment. If Pacifica were to devote all its resources to manufacturing, ZM units of labor would be unemployed.

Before the arrival of the development team all resources were fully employed in subsistence agriculture.[23] The availability of new technologies produces the factor proportions problem, which occurs when the only available production techniques use factors in a different proportion than that in which factors are available. Consequently, one or more factors are redundant, or unemployed. Furthermore, if labor is unemployed at M_4, reductions in wages will not cause an increase in employment. A Keynesian solution will not eliminate unemployment either. Since the capital stock is limited and the way in which labor can be combined with it is also limited, further increases in demand would simply be inflationary. The economy is already on the transformation curve and no further increases in real output can occur.[24]

Suppose the development advisers from the United States are sent to help plan

---

[23] This assumes that the capital stock before the technological assistance program was only large enough to support subsistence agriculture, implying that foreign aid increased the width of the box by an amount OA_L.

[24] Keynesian remedies are only applicable when the economy is inside the transformation curve. In that case, employment and output can be increased in some industries without reducing output elsewhere.

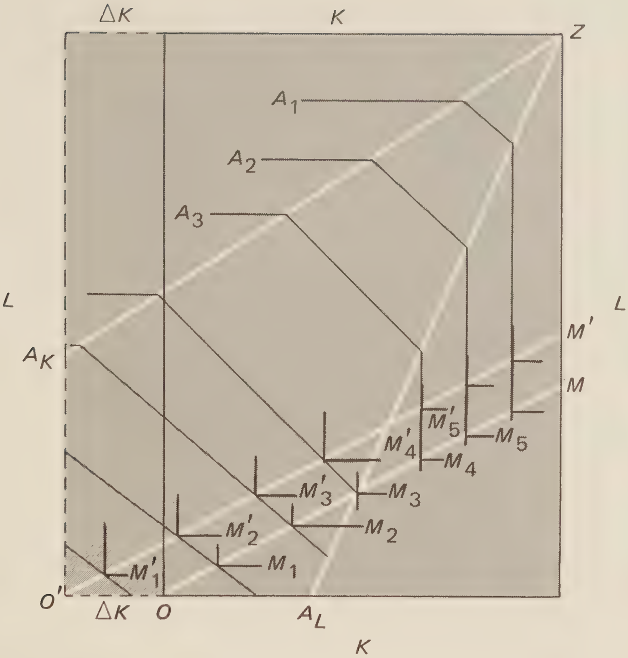


FIGURE 6–21
Impact of an Increase in Capital

the use of some foreign aid that will be used to purchase capital. Then the *K* sides of the Edgeworth box are extended, as shown in Figure 6–21. The origin of the manufacturing isoquants is shifted from *O* to *O'*. Before the increase in *K*, at manufacturing output M_4 there was unemployment of labor. After the increase, labor is fully employed at M_4. However, at M_5 there is a factor proportions problem again. A much larger amount of capital would be required to eliminate the possibility of unemployment for any level of manufacturing output.

Say that a new process is discovered for producing manufactured goods that is more labor-intensive than *M*. In Figure 6–22 the original process is labeled M_K and the new process M_L. The efficiency locus is *OEZ*. That is, as manufacturing output increases, process M_L and a combination of A_K and A_L are used until point *E* is reached. After *E*, process A_K and a combination of M_L and M_K are used. In no case is there a factor redundancy. All points on the transformation curve are associated with full employment of both *K* and *L*. While this does not necessarily happen in all cases, it does show that an alternative to increasing capital stock is simply to use capital more sparingly in production. If unemployment is due to a factor proportions problem, finding more labor-intensive production processes is an alternative to increasing the rate of investment.[25]

[25] The impact of these alternative development strategies on labor productivity will be different, however. Labor-intensive processes are associated with lower marginal labor productivity and hence lower wages. The strategies will depend upon the relative weights given to alternative development objectives.

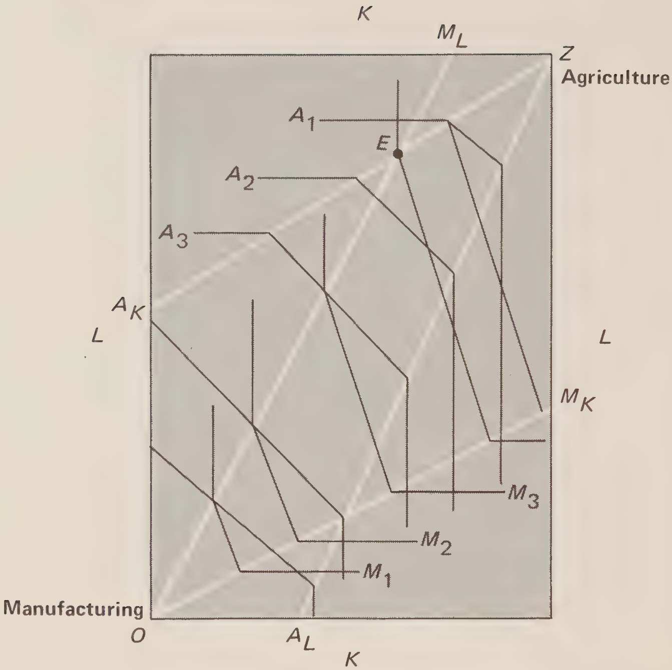


FIGURE 6–22
Impact of a Change in Technology

6–4 PROFIT MAXIMIZATION AND THE SUPPLY CURVE FOR THE FIRM

Up to this point we have examined the link between technology, factor availability, and production costs. Given a cost function or transformation curve, the individual firm must determine the optimal scale of operations. To analyze this problem we must take into account the objectives of the firm and thus reenter the institutional context within which production takes place.

Regardless of whether the firm operates in a market system or in a centrally planned economy, the optimal scale of operation will depend upon demand. An enterprise in a market system will want to maximize profits or sales, both of which depend on demand. In a socialist context the central planning board attempts to set prices to ensure that everything produced will be purchased and that there is no excess demand. Thus nonmarket prices are also governed by demand conditions.

Consider a small firm operating in a market system. The firm is sufficiently small in relation to the total market that its operations have a negligible effect on market supply. The firm is thus unable to influence market price by its operations and views market price as given or fixed. It can presumably sell all it can produce at that price.[26]

[26] We will drop this assumption in chapter 9 when we consider alternative models of market structure.

We assume that the firm's objective is to maximize profits, a reasonable assumption about a small firm where the owner is also likely to be the manager who sets priorities.[27]

Profit is defined as total revenue minus total cost, that is,

$$\pi = TR - TC$$

where π is profit. Profits can be increased as the scale of operation increases as long as the increment to total cost is less than the gain in total revenue, that is,

$$\frac{\Delta \pi}{\Delta X} > 0$$

when

$$\frac{\Delta TR}{\Delta X} > \frac{\Delta TC}{\Delta X}$$

Since

$$\frac{\Delta TR}{\Delta X} = p$$

and

$$\frac{\Delta TC}{\Delta X} = MC$$

this implies that maximum profit occurs where price is equal to marginal cost when marginal cost is increasing, that is,

$$p = MC$$

for

$$\frac{\Delta MC}{\Delta X} > 0$$

The relation is shown graphically in Figure 6–23. *AC* and *MC* are assumed to be U-shaped. Notice that this assumption is absolutely essential to the notion of profit maximization, since if *MC* were falling or constant over the entire range of production possibilities, the firm could expand indefinitely without ever reaching its maximum profit.[28]

When price is p_1, output will be X_1. If demand increases, causing price to rise

---

[27] Some managers may have other objectives. One commonly suggested is the maximization of sales, particularly when managers who are not owners of the firm are responsible for establishing priorities. Sales maximization is incompatible with the assumption that firms can sell all they can produce at a fixed price, however. The sales maximization solution would require a firm to expand indefinitely, after which it would become large enough to affect market price. We will consider this model in chapter 9, where we drop the assumption that a firm is small in relation to the market.

[28] If the firm becomes so large that it can affect market price, at large scales of output market price would eventually fall below *MC* even if *MC* is constant or decreasing.

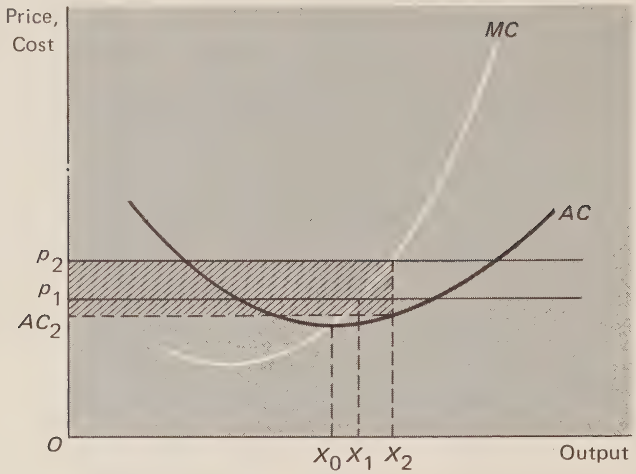


FIGURE 6–23
Profit Maximization Market Model

to p_2, output will increase to X_2. Profit per unit is the difference between selling price and unit cost. The shaded area represents total profit when price is p_2.

If *AC* is greater than price, the firm will be operating at a loss (negative profit). Since price is equal to *MC*, and since *MC* intersects *AC* at its lowest value, this implies that the firm will never operate in the decreasing range of its *AC* curve.

The marginal cost curve to the right of X_0 is the **supply curve** for the firm. The supply curve relates quantity supplied to market price. For any market price, the quantity demanded can be read from the *MC* curve, provided *MC* is greater than *AC*. The market supply curve is the horizontal sum of the individual supply curves of the firms in the market.

If we have a model of behavior in a nonmarket economy, we can also determine the optimal scale of operation for these firms. In the Lange model, firms are required to combine inputs in such a way that the cost of production is minimized for every level of output. As we have seen, this requires that the marginal rate of factor substitution equal the ratio of factor prices, that is,

$$\frac{p_K}{p_L} = \frac{MP_K}{MP_L} = MRS_{LK}$$

Thus socialist managers in the Lange model will select exactly the same expansion paths as their counterparts in a capitalist system.[29]

Furthermore, in a Lange system firms are required to expand production to the point where marginal cost is equal to the price established by a central

[29] This is only true if relative factor prices and production technology are the same in both institutional contexts. Differences in factor prices (given similar technologies) could be due to different conditions of factor supply associated with differences in ideology of the respective societies. Differences in technology are likely to be negligible if we assume equivalent levels of development.

planning board. This is true regardless of the size of the firm, since the price is fixed by the CPB and presumably cannot be influenced by the firm's market activities, regardless of the scale of its operations. Thus Lange firms will also select the same scale of operations as their capitalist counterparts.

Since Lange firms are not directly concerned with profits per se (although Lange specifically states that his managers perform the same function that capitalist managers perform when they maximize profits)[30] the relation between average and marginal cost does not matter to them. Consequently, the entire *MC* curve is the supply curve of a Lange firm. If the *MC* curve is U-shaped, intersecting the price line at two points, the firm is expected to select the larger output and will produce in the upward sloping range of the *MC* curve. If *MC* is decreasing over the entire range, marginal cost pricing will actually produce a minimum profit position. But since profits are not a factor in the production decision, Lange firms will set marginal cost equal to price over the entire range of the *MC* curve.[31] Despite the fact that profits are being minimized when *MC* is decreasing, the marginal cost criterion ensures that goods will be produced at equal marginal rates of transformation throughout the economy, which is necessary to achieve Pareto efficiency.[32]

6–5 TECHNOLOGICAL CHANGE

We have assumed that the firm makes its production decisions on the basis of a fixed production function. Given technology and factor and product market conditions, the firm determines the optimal factor mix and scale of production.

However, the firm may attempt to alter its profit picture by changing its production function, that is, it may actively seek more efficient ways to produce its output. Technological change involves finding new production techniques not previously available. Thus it is important to distinguish between a change in the production function and simply a movement to a new process, which was previously available, in response to changing factor prices or availabilities.

Firms will seek to improve their production function so that they can produce

---

[30] Oskar Lange, *On the Economic Theory of Socialism* (Minneapolis: University of Minnesota Press, 1938), p. 76.

[31] Lange never addresses himself to the question of decreasing marginal cost. He consistently asserts that his model will produce behavior similar to that of a competitive market system. As we have seen, firms in a market system (facing fixed selling prices) will never produce in the falling range of their *MC* curves. We will consider the impact of economies of scale on capitalist production decisions in chapter 9.

[32] If *MC* falls fast enough to produce a transformation curve that is convex to the origin, then the condition

$$MRS = MRT$$

may produce a minimum rather than maximum welfare condition. For a discussion see Francis M. Bator, "The Simple Analytics of Welfare Maximization," *American Economic Review, XLVII* (March 1957), pp. 43–52.

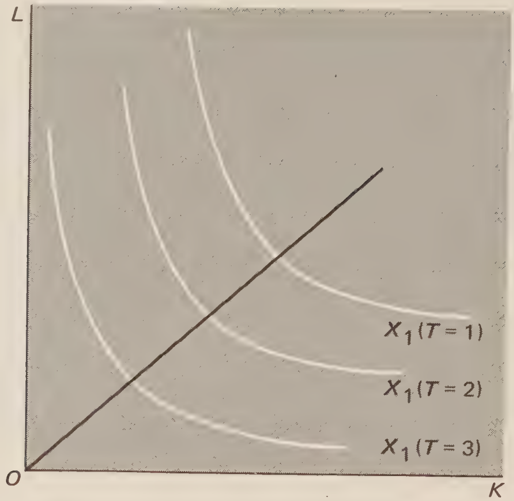


FIGURE 6–24
Neutral Technological Change

the same output with fewer inputs. Technological change can affect the production function in a number of ways. Suppose technological change occurs over three time periods, $T = 1$ to $T = 3$. If it shifts the isoquants toward the origin, as shown in Figure 6–24, but leaves their shape unchanged, technological change is said to be **neutral.** If factor prices are fixed, the optimal expansion path is unaffected and factor proportions remain the same. However, factor productivity increases for each factor.

Technological change can affect optimal factor proportions. In Figure 6–25

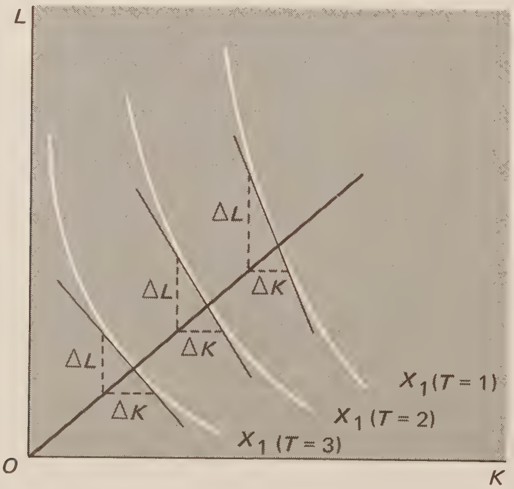


FIGURE 6–25
Labor-Using Technological Change

the technological change causes the isoquants to shift toward the K axis. If factor proportions were constant, the marginal rate of substitution of labor for capital will fall from $T = 1$ to $T = 3$, implying that the marginal productivity of capital decreases relative to the marginal productivity of labor. If the factor prices remain fixed, the optimal expansion path will become more labor-intensive from $T = 1$ to $T = 3$. Thus, since it induces firms to employ more labor per unit of capital, this type of technological change is called **labor-using.** If technological change causes the isoquants to become steeper, the change is called **capital-using** for analogous reasons. Capital-using technological change raises the marginal productivity of capital relative to the marginal productivity of labor.

Now it is clear that we must make a distinction between a change in technique and technological change. If technological change is neutral, there will be no change in technique, although factor productivity has increased. However, costs will be lower for each level of output, so that the scale of output may be affected if the firm is maximizing profits.[33] If the production function is not homogeneous, the scale effect may cause the firm to alter factor proportions.

If technological change is capital-using or labor-using, it affects technique as well as efficiency. Capital-using technological change causes production techniques to become more capital-intensive, increasing capital productivity relative to labor productivity for the former capital-labor ratio. Firms operating in an environment in which labor is relatively scarce will seek to adopt capital-using technological change. In many cases, we export such technological "improvements" to our less-developed neighbors, who are plagued with a scarcity of capital and abundance of labor. Clearly, by increasing the capital requirement per unit of labor input, such "improvements" can only intensify the factor proportions problem there. Technological innovation generally requires rather sophisticated familiarity with modern industrial techniques. Thus it is likely that technological advance in most industries is likely to be developed in industrially mature economies, so that the bias will be toward the capital-using variety.[34]

Technological change has the desirable effect of reducing production costs, which may result in reduced product prices or improved quality. However, technological change also affects the demand for factor inputs by the firm as well as the way in which income is distributed among the factors of production. Much

---

[33] The effect on profits and the scale of output for the individual firm will obviously depend on the extent to which the technological change is diffused throughout the industry. If many firms adopt the change, industry output will expand concomitantly, causing market price to fall by more than if a single firm had initiated the change. This explains why firms jealously guard their research plans and discoveries. Since research is potentially more profitable if other firms are barred from access to new inventions, the patent system, which limits the diffusion of inventions, is said to encourage research and product improvement.

[34] There is evidence, however, that technological change in the United States has become labor-using in recent years. See C. E. Ferguson, "Substitution, Technical Progress, and Returns to Scale," *American Economic Review, Papers and Proceedings, LV* (May 1965), pp. 296–305. This implies that firms in the United States view capital rather than labor as the limiting factor of production. Nevertheless American technology is still capital-using with respect to technology available in the underdeveloped world.

of the structural unemployment in the American economy today is attributed to technological change. In chapter 7 we will consider the impact of technological change on factor prices and the demand for factor services.

Technological Change and the Objectives of Firms

Technological change is one option available to firms for increasing profits. Other options include advertising, which will increase demand for the product, and product improvement, which will capture a larger share of an existing market from competitors. Presumably, a profit-maximizing firm will engage in any of these activities as long as the addition to total revenue exceeds the addition to cost.[35]

However, the question of the optimal rate of technological change is somewhat more complicated. To begin with, there is a long and uncertain lag between the initiation of research and development and the creation of a usable invention or innovation. Thus the actual payoff of any project is extremely uncertain. Most economists agree that there are substantial economies of scale to research and development (R and D) activities, so that only large-scale firms can afford to undertake such projects. Small firms undertake technological change only after knowledge is diffused throughout the industry and is no longer protected by patents. The patent system is often thought to benefit larger firms, who can afford R and D, and to give them a competitive advantage over small firms. On the other hand, it is argued that the patent system is necessary to encourage research and development.

Galbraith has argued that firms must be large and, in fact, dominate the market if technological change is to proceed in that industry.[36] Other economists like Mansfield have found that large firms spend more for R and D (as a proportion of sales) than small firms but that very large firms spend no more than moderately large firms.[37] This conclusion supports the hypothesis that there are economies of scale in innovative activities that lead to technological change but that mammoth corporations are not required to take advantage of them. It is also likely that the size at which firms can take advantage of economies of scale in R and D will vary considerably with the type of product and new processes developed. If new processes are available only in very large units, then clearly they will be advantageous only for large firms. We shall return to the question of the relation between market structure and technological change in chapter 9.

[35] The question of developing appropriate incentives for technological change and product improvement in an economic system without profit incentives is one of the thorniest problems of socialist economies. In many noncapitalist systems there is a growing pressure for the use of profit incentives as opposed to marginal cost pricing. Similar difficulties arise in capitalist economies, where certain public utilities are regulated by a system of marginal cost pricing.

[36] John Kenneth Galbraith, *American Capitalism* (Boston: Houghton Mifflin, 1962), pp. 84–94.

[37] Edwin Mansfield, *Technological Change* (New York: W. W. Norton, 1971), pp. 68–71.

QUESTIONS FOR STUDY AND REVIEW

1. A bootlegger can make three types of whiskey, which he can sell on the black market at $3 per gallon. He has 24 stills and 16 retorts available. To produce sour-mash bourbon he needs 1 still and 4.5 retorts per gallon; for white lightning he needs 2 stills and 3 retorts per gallon; and rot-gut gin requires 4 stills and 1 retort per gallon. Each process takes one day.
   a) What is his optimal operating plan?
   b) Would he rent an additional still or retort for 75 cents per day?
2. An economy uses two factors, K and L, to produce two commodities, A and B. The available supply is 150 units of K and 100 units of L. Commodity A can be produced by two alternative processes:

   Process A_1: One unit of A requires 3 units of K and 3 units of L.
   Process A_2: One unit of A requires 5 units of K and 2 units of L.

   Commodity B can be produced by one process only:

   Process B_1: One unit of B requires 8 units of K and 4 units of L.

   Assume commodity prices, p_A and p_B can take on any value, depending on demand conditions, but the economy maximizes revenue at whatever set of prices are established.
   a) Construct the efficiency locus in an Edgeworth box diagram.
   b) Construct the transformation curve.
   c) Can you determine one or several commodity price ratios, p_A/p_B, at which unemployment of either factor will occur?
3. In the economy described in Question 2 a new process is discovered for producing B that requires 4 units of K and 3 units of L. What is the effect on the transformation curve? Can unemployment occur now? Explain.
4. "Although costs must be minimized for profits to be maximized, the reverse is not the case." Carefully explain this statement, describing the necessary conditions for cost minimization and profit maximization.
5. Compare the neoclassical analysis of production with the neoclassical analysis of household behavior, pointing out both similarities and differences.
6. Discuss the implications of linear homogeneity for
   a) Optimal factor proportions.
   b) Short-run average cost (some factors fixed).
   c) Long-run average cost (no factors fixed).

   Do you find linear homogeneity a reasonable assumption for empirical work? Be sure to justify your answer.
7. One of the goals of economic development is to absorb labor from subsistence agriculture into the modern economy. From your knowledge of the theory of production, what development strategies would you recommend?

ADDITIONAL READING

Baumol, William J. *Economic Theory and Operations Analysis,* 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1972, chapter 12.

Bronfenbrenner, Martin, and Paul Douglas. "Cross Section Studies in the Cobb-Douglas Function," *Journal of Political Economy, XLVII* (December 1939), pp. 761–783.

Chamberlin, Edward H. *The Theory of Monopolistic Competition,* 8th ed. Cambridge, Mass.: Harvard University Press, 1962, appendix B.

Dorfman, Robert. "Mathematical, or 'Linear,' Programming: A Nonmathematical Exposition," *American Economic Review, LXIII* (December 1953), pp. 797–825.

Eckhaus, Richard S. "The Factor Proportions Problem in Underdeveloped Areas," *American Economic Review, L* (September 1955), pp. 539–565.

Ellis, Howard, and William Fellner. "External Economics and Diseconomies," *American Economic Review, XXXIII* (September 1943), pp. 493–511.

Leibenstein, Harvey. "The Proportionality Controversy and the Theory of Production," *Quarterly Journal of Economics, LXIV* (November 1955), pp. 619–625.

Mansfield, Edwin. *Technological Change.* New York: W. W. Norton, 1971.

Ruggles, Nancy. "The Welfare Basis of the Marginal Cost Pricing Principle," *Review of Economic Studies, XVII* (1949–50), pp. 29–46.

Viner, Jacob. "Cost Curves and Supply Curves," *Zeitschrift für Nationalökonomie, III* (September 1931), pp. 23–46. Reprinted in American Economic Association. *Readings in Price Theory.* Chicago: Richard D. Irwin, 1952, pp. 198–232.

7 Productivity and Distribution

The mirror image of the theory of production is the theory of resource utilization and factor price determination. The distribution of resources among individuals in a society, the rate of utilization, and the payment for the use of these resources in production determines the distribution of income.

Since values concerning the optimal distribution of income vary considerably among societies, it is surprising to learn that the principles governing the distribution of income in production are the same regardless of the institutional arrangements. Income inequality may be due to different tax structures and inheritance laws; differences in the right to acquire land, capital, and property; and differences in the opportunity to acquire education and other experience conducive to high productivity in the labor force. Regardless of how resources are distributed, however, the conditions under which they are employed, that is, their remuneration and rate of utilization, can be derived from the theory of production, which is common to all economic systems.

7–1 THE MARGINAL PRODUCTIVITY PRINCIPLE

Suppose an increase in demand for a product causes the market price (or the price set by a central board) to rise. The optimal level of output will increase, and the firm will immediately seek to acquire new resources. If additional planning and delays in acquisition would be required to expand the capital stock, the firm could increase output in the short run only by hiring additional labor.

In the market model the profit-maximizing firm faces a fixed market price for goods and resources. The acquisition of additional units of labor will increase profits as long as the addition to revenue, the marginal product multiplied by the product price, is greater than the addition to cost, the wage. Assuming diminishing returns to labor when capital is fixed, MP_L declines as more labor is hired, and the firm will hire labor until

$$p \times MP_L = w$$

Marginal product multiplied by the product price is called the **marginal value product.** The marginal value product schedule is the demand curve for labor. It will have the same slope as the MP_L schedule. Given the wage, the quantity of labor employed by the firm can be determined from the *MVP* schedule, as shown in Figure 7–1.

Suppose the *MVP* schedule is MVP_0 and the wage is w_0. L_0 units of labor will be employed. Then an increase in the product price occurs that produces a vertical upward shift in the *MVP* schedule equal to the change in price. That is,

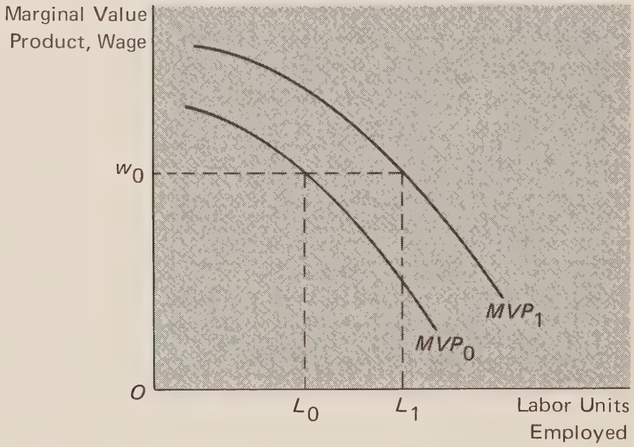


FIGURE 7–1
The Short-Run Demand for Labor

at each level of employment *MVP* increases by the amount of price increase. In Figure 7–1 employment increases from L_0 to L_1 when the *MVP* schedule shifts from MVP_0 to MVP_1.

Consider a firm operating in a Lange economic system, setting marginal cost equal to price. The CPB raises the product price in response to an excessive demand. If labor is the only variable factor of production in the short run, marginal cost is the wage divided by the marginal product, that is,

$$MC = \frac{w}{MP_L}$$

If MP_L is 2 units per hour and the hourly wage is \$1, the marginal cost per unit is 50 cents. The firm will hire additional units of labor until

$$p = MC = \frac{w}{MP_L}$$

or

$$p \times MP_L = w$$

Thus managers of Lange enterprises behave in exactly the same way as profit-maximizing capitalists. The demand for labor is determined by its marginal value product.

Now, if the firm begins to acquire more labor with other factors held constant, it will begin to deviate from the optimal expansion path. As employment of labor increases in response to wage cuts, the firm will eventually want to acquire more of the other factors of production. But when more nonlabor factors are employed, MP_L rises for a fixed amount of L, causing the MP_L schedule to shift to the right. This means that the firm's actual long-run demand curve for labor is somewhat flatter than the MVP schedule.

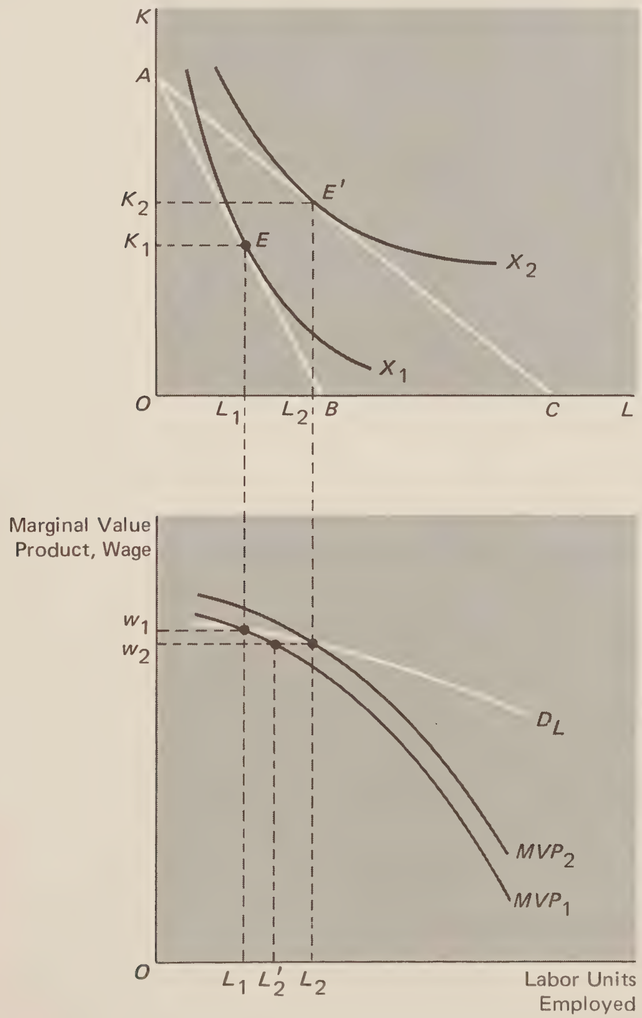


FIGURE 7-2
Impact of a Wage Decrease on Employment

Suppose there is a decline in the wage. If labor is the only variable factor of production, the impact on employment can be found from the *MVP* schedule. However, if all factors of production are variable, so that the firm can move along the optimal expansion path, the results are like those in Figure 7–2: A decline in the wage causes the cost line to shift from *AB* to *AC*, and the amount of labor employed rises from L_1 to L_2 and capital from K_1 to K_2.[1]

The increase in *K* causes the *MVP* schedule for labor to shift from MVP_1 to MVP_2. If *K* had remained fixed, the demand for labor would only have been L_2'. The actual demand curve for labor, D_L, is flatter than the *MVP* curve if the firm acquires additional capital when wages fall. Notice, however, that in equilibrium the wage is always equal to the marginal value product of labor, even though *MVP* may be affected by the change in other factor inputs.

The marginal productivity principle of distribution states that factor payments will be proportional to the marginal productivity of the factor. The constant of proportionality is the product price. The marginal productivity principle implies that the factor payment reflects the value of the last unit of that factor in production. People who receive low wages are less productive (at the margin) than those who receive high wages. If a low-paid worker quits, the impact on the total value of output is less than if a high-paid worker leaves the labor force. If I move to a higher paying job, the value of total output should increase.

The Market Demand Curve

The market demand curve for a factor of production is the horizontal sum of the demand curves of individual firms. Suppose the going wage for pipemakers is established at $1.50 per hour. A minimum wage law is enacted that raises the wage to $2.00. What will be the impact on the demand for pipemakers?

If we add up all the labor demand curves for all firms that employ pipemakers, we obtain the schedule labeled Σd_1 in Figure 7–3. Suppose the wage rises from w_1 to w_2. As employment declines, industry output must decline also. Now when a single firm contracts its output, the effect on the product market price is negligible. But if all firms in the industry contract output, the reduction in market supply will cause price to rise. This will have the impact of shifting the *MVP* schedules for all firms in the industry, so that their aggregated labor demand schedule rises to Σd_2. Employment of labor at w_2 declines to L_2 rather than L_2'. Thus the market demand curve for labor, D_M, is steeper than either Σd_1 or Σd_2, since output price adjustments in response to changes in product market supply must be taken into account when analyzing aggregate market demand for a factor.

The impact of output changes on market price will depend upon the elasticity

---

[1] The isoquant map is used to locate the new *K/L* expansion path, but *E'* is not the final equilibrium for the firm. Since *E'* entails the same level of total costs as *E* at the higher wage, while output has increased, the total cost curve shifts to the right. Unit costs will decline for all levels of output, so that the firm will expand its scale of operation. If the production function is homogeneous, factor proportions will remain unaffected and the firm will expand along the expansion path through *E'*.

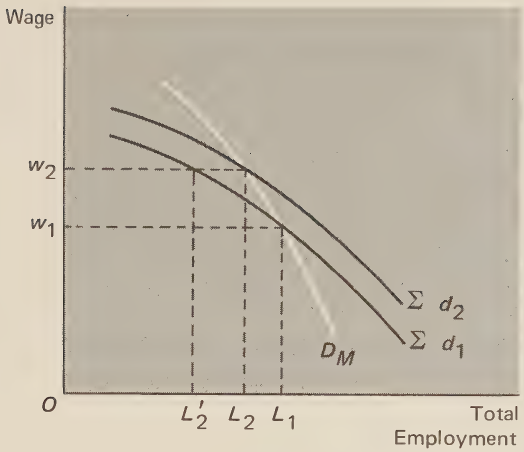


FIGURE 7-3
Derivation of the Market Demand Curve for a Factor

of demand for the product. The less elastic the demand for the product, the less elastic the demand for the factors that produce it and the steeper will be the market demand curve for the factor in relation to the aggregated *MVP* schedule. Notice, however, that at all points on D_M the marginal productivity principle applies. Wages are equal to marginal value product.

The Impact of Technological Change

Technological change may have the effect of raising the marginal productivity of labor and other factors. Different types of technological change will affect the MP_L schedule differently. For instance, labor-using technological change raises the marginal product of labor relative to capital more than neutral or capital-using technological change. Nevertheless, unless they are extremely capital-using, all types of technological change will raise the marginal productivity of labor to some degree.

If technological change is limited to a single firm and if the firm is so small that changes in its output have no impact on market price, then a rightward shift in MP_L will shift *MVP* proportionately and the demand for labor will increase. However, technological change is likely to be initiated by large firms whose operations have an impact on market price. In many cases, it is initiated by one firm and adopted by the whole industry. If product market demand is inelastic, the impact on price may be so great that *MVP* actually decreases. In that case, technological change would be causing a decline in the demand for labor. The direction of the change in the demand for labor will depend upon the factor bias of the technological change, the magnitude of the increase in overall factor productivity, and the elasticity of product demand.

Several years ago many tire companies introduced automated assembly line techniques. Automated techniques are highly capital-using, and the demand for tires is price inelastic (the demand for tires may be more sensitive to automobile prices than tire prices). The Rubber Workers Union vigorously protested the change, arguing that it would displace many workers, as in fact, it did. There was a substantial increase in overall factor productivity, but labor productivity rose relatively little. The industry is dominated by four major companies and the new techniques were adopted by all four. If output had been increased, profits would have declined because of inelastic demand. Consequently, the firms maintained their previous outputs while technological change allowed them to reduce the amount of labor required to produce them. Since the change was capital-using, the major impact was on labor, and the level of employment declined.

7–2 FACTOR SUPPLY AND DISTRIBUTION

Since remuneration depends not on a worker's actual contribution to production, his *average* productivity, but on the contribution of the last worker hired, or *marginal* productivity, it is important to examine the determinants of labor supply. Clearly, if I work in an occupation for which there are many eager applicants willing to work for a low wage, the marginal product may be considerably lower than in an occupation in which no one wants to work.

In chapter 5 we examined the determinants of factor supply. Households will offer productive services to the market until the marginal rate of substitution of income for leisure is equal to the wage, that is,

$$MRS_{Yl} = w$$

An increase in the wage will cause an increase in labor supplied as long as the substitution effect dominates the income effect of the work-leisure choice. This analysis examines the determinants of the choice between work and leisure. Thus it can explain why people choose to enter the labor force. It does not explain, however, why conditions of supply vary among occupations. In short, it is a theory of labor force participation, not a theory of occupational choice.

It is a commonplace observation that conditions of supply vary markedly among occupations. The relation between the supply of nurses and their wages is different from that between the supply of doctors and their wages, and supply in both of these occupations differs markedly from the supply of airline pilots. There are a number of reasons why supply functions, the relation between the wage and the labor supply, differ among occupations. Three major ones are psychic rewards or income on the job, training costs, and need for personal attributes that are in limited supply.

Clearly, some jobs are less disagreeable than others. Economic theory has traditionally focused on product consumption as the sole source of utility or economic satisfaction. But some jobs provide a source of psychic gratification

over and above the provision of income to buy goods and services. To the extent that there is consumption on the job, the money wage does not reflect the true remuneration. Teaching and nursing are examples of jobs that are low-paying relative to the skill required but that entail considerable psychic rewards. A lawyer who leaves his practice for a law school faculty may accept a substantial cut in his money income to escape the frustrations of dealing with people's problems and to gain the psychic benefits associated with teaching. Thus the greater the psychic reward, the greater the supply of labor associated with any wage for that occupation. To attract labor into more disagreeable jobs the wage must be increased.

A second determinant of differences in supply is training costs. Suppose I find high school and college teaching equally attractive occupations. However, to teach in a university I need a Ph.D. degree, while a bachelor's degree is sufficient for high school teaching. In order to attract me into college teaching the wages there relative to high school teaching must be sufficiently high to remunerate me for the cost of obtaining a Ph.D. degree after college. These training costs include the earnings I forgo by remaining out of the labor force as well as tuition and other educational expenses.

Some economists have attempted to compare the costs of education with wage differentials attributable to higher education.[2] Gary Becker[3] has found, for instance, that the typical male college graduate earns about $100,000 more over his lifetime than a typical high-school graduate.[4] Lester Thurow has estimated that a college education increases the yearly income for the average male by almost $3,000.[5] However, such studies usually do not take into account the possibility that college graduates on the average hold more agreeable jobs than high school graduates. If this is true, then such studies may understate the true difference in remuneration.

If psychic rewards are higher in jobs requiring more skill, it would seem that wage differentials should be less than training costs. In fact, empirical evidence suggests that the reverse is the case, that the more training is required the greater is the differential that cannot be attributed to training costs. This suggests that differences in ability or the opportunity to acquire training restrict the supply of labor in highly skilled professions. Workers in those occupations can earn wage differentials in excess of training costs, because new workers cannot be attracted into the market by monetary incentives. Such differentials are called **rents.** We will analyze the determinants of rent later in this chapter.

In Figure 7–4 we can see how the wage in any labor market is determined

[2] A good collection of these studies is Mark Blaug (ed.), *Economics of Education: 1* (Baltimore: Penguin Books, 1968).

[3] Gary Becker, *Human Capital* (New York: Columbia University Press, 1964).

[4] To compare these earnings with the costs of education they must be discounted by the market rate of interest. That is, if the money used for education were invested at interest, the total lifetime earnings would be greater than the initial investment. See chapter 10 for further discussion of this point.

[5] Lester Thurow, *Investment in Human Capital* (Belmont, Calif.: Wadsworth Publishing Co., 1970), p. 22.

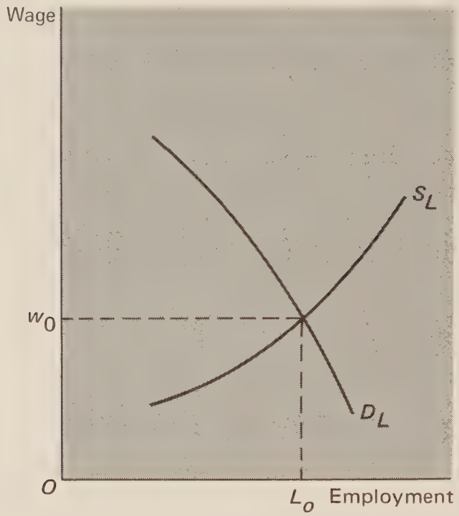


FIGURE 7–4
Wage Determination in a Labor Market

by the interaction of market supply and demand. Once this market wage is established, individual firms acquire labor according to the marginal productivity principle.

Analysis of Occupational Wage Differentials

We have examined the determinants of the market demand curve for labor in an occupation, and we have explored some factors influencing labor supply and occupational differences in labor supply. One of the most interesting applications of the theory of wage determination is the explanation of wage differentials among occupations. Wage differentials can be attributed to differences in conditions of demand or supply, or both. Assume that the marginal value product schedules for doctors and nurses are the same. The supply functions are different because of the higher training costs for doctors. In Figure 7–5 the wage for doctors, w_D, is greater than that for nurses, w_N. If there were no barriers to entry into the physician category and no psychic benefits accruing to one profession over the other, the differential should reflect training costs alone.

There are, however, other factors besides training costs that limit the supply of doctors. The ability and opportunity to pursue medical training is not available to everyone. For instance, the medical profession notoriously discriminates against women. Thus most nurses would not be able to become physicians even if w_D were greatly in excess of w_N. Consequently, a differential may be maintained in excess of the difference in training costs.

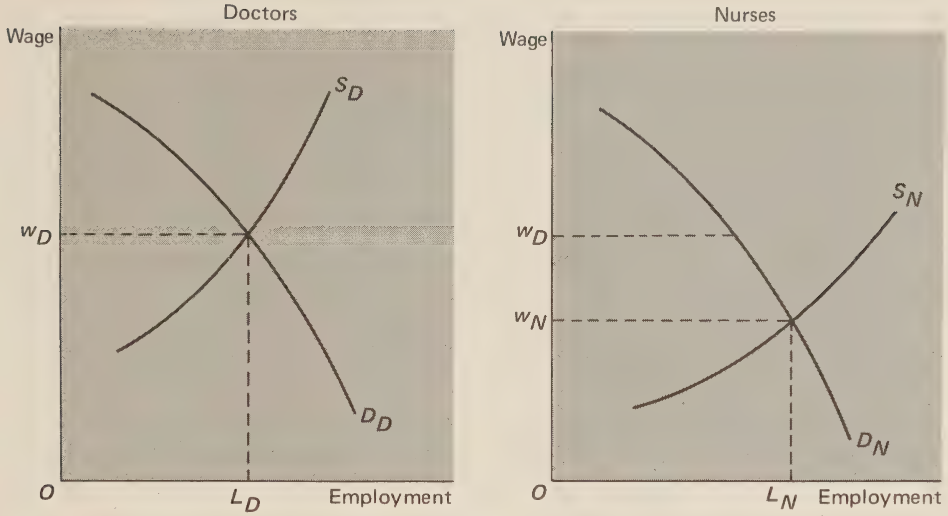


FIGURE 7–5
Wage Differentials Due to Supply Conditions

Wage differentials may also be due to differences in demand. Marginal productivity for any level of labor input will be greater the more capital-intensive the industry. Therefore, wages will be higher in the steel industry than in the textile industry if supply conditions are the same for both. In Figure 7–6 the *MVP* schedule for steel workers is above the *MVP* schedule for textile workers, although the supply functions are the same. Wages in the steel industry, w_S, are higher than in the textile industry, w_T. If labor is mobile between these industries, the higher wage in steel should induce workers to leave textiles. However, there are many reasons why labor cannot move between industries in response to wage differentials.

There may be geographic specialization of industry, and workers will generally not move to another part of the country to earn higher wages. Specialized skills and experience may also prevent interindustry mobility of labor. In addition, discrimination against certain groups in the labor force in high-productivity industries may perpetuate wage differences. The textile industry in the United States has traditionally employed women and ethnic minority groups, while the more capital-intensive industries often restrict entry of these groups and favor white native-born males.[6]

In most cases, wage differentials are due to differences in both demand and supply. Many high-productivity jobs also require considerable training and experience. It is likely, for instance, that *MVP* for doctors is greater than that for nurses. Doctors are trained to do many more complicated tasks than nurses can

[6] All these factors may be viewed as restrictions in labor supply to the steel industry. Using this interpretation, the assumption of identical supply functions is no longer valid.

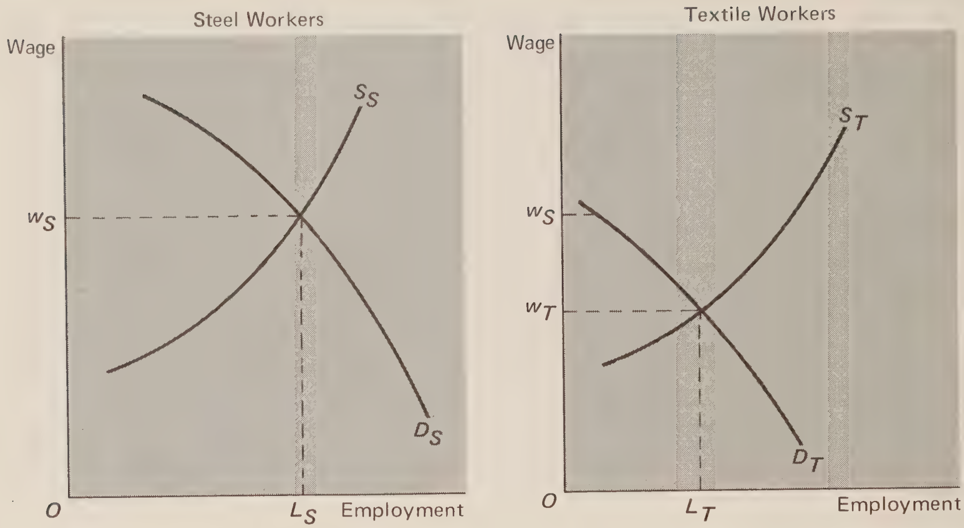


FIGURE 7–6
Wage Differentials Due to Demand Conditions

perform. If this is the case, the wage differential between doctors and nurses is attributable to differences in both demand and supply. In Figure 7–7 the demand for doctors is shown to be greater than that for nurses. If demand were the only consideration, the wage differential would be $w_D - w_N'$. The actual differential $w_D - w_N$ is due to the additional effect $w_N' - w_N$ of a lower supply curve for nurses.

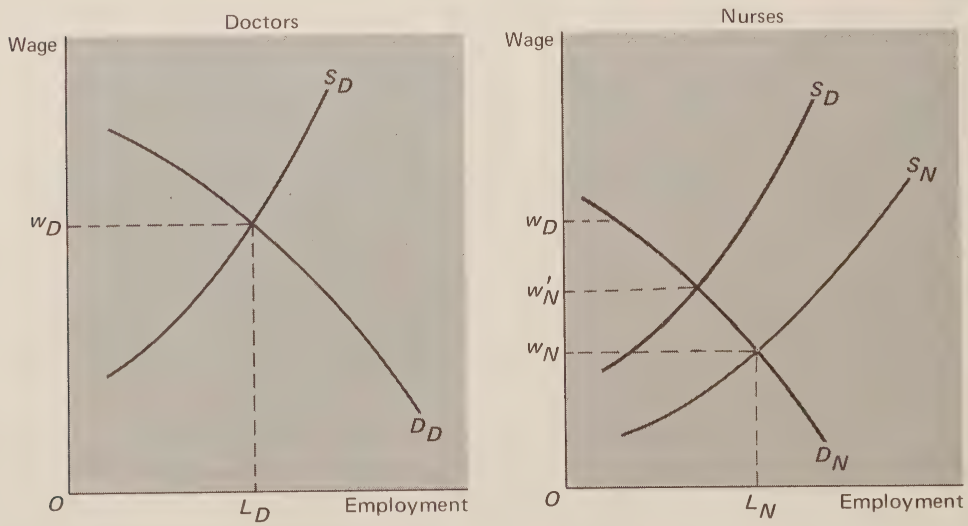


FIGURE 7–7
Wage Differentials Due to Demand and Supply Conditions

Wage Discrimination

Discrimination occurs when differentials in wages are associated with personal characteristics of the worker that are not directly related to productivity. Blacks and women as groups earn considerably less than their white male counterparts.

Many economists attribute discrimination to differences in supply associated with the acquisition of appropriate skills. Blacks have been excluded from acquiring the same educational opportunities as whites, and cultural factors have prevented them from obtaining the appropriate modes of behavior requisite to satisfactory performance in a white working environment, particularly high-status jobs requiring substantial personal interaction. Thus blacks have a lower supply price than whites and in addition will have to seek jobs requiring fewer skills, which are likely to be low-productivity jobs.

Discrimination may also take the form of excluding certain groups from high-productivity jobs. Although white women have better educational opportunities than blacks, they are often excluded from high-productivity jobs on the basis that such operations are not "women's jobs." Sex role identification extends from the family to the marketplace. The most "appropriate" jobs for women are those associated with her household activities—nurse, teacher, interior decorator, secretary (man's helper), and so on. Thus exclusion of women from high-productivity jobs also explains persistent wage differentials on the basis of sex.

Discrimination is not necessarily inconsistent with the marginal productivity principle. In many cases, blacks and women are perceived by employers to be less productive than white men with equivalent educational and professional experience. This is most likely to be the case in service industries where the recipients of the services prefer certain types of workers—men, whites, persons over 25—and where "output" is difficult to measure objectively. It is also likely to occur in white-collar professions where the preferences of other employees are an important consideration. Hiring a black or a woman to supervise an office of white males might be expected to reduce the productivity of the entire office.

Many economists have argued that overemphasis on marginal productivity has obscured the analysis of race and sex differences in wages. They suggest that a more fruitful way to approach the problem is to examine why differences in perceived (and real) productivity persist. Barriers to the attainment of human capital and appropriate attitudes toward work, as well as social attitudes toward women and ethnic minorities, need to be examined if we are to understand the real basis for discrimination.

7-3 THE THEORY OF RENT

We have seen that wage differentials in excess of differences in training costs may persist among occupations. If this is true, labor in the favored occupations

is earning a **rent.** Rent is the difference between the actual wage and the hypothetical wage required to attract labor into that occupation.[7]

Suppose doctors and nurses are equally productive and the professions have the same psychic rewards. In Figure 7–8, S_D is the supply schedule for doctors and S_N is the supply schedule for nurses. The wage differential will be *AB*. Suppose *AC* is the differential required to make up for differences in training cost. Then *CB* is a differential rent occurring to physicians.

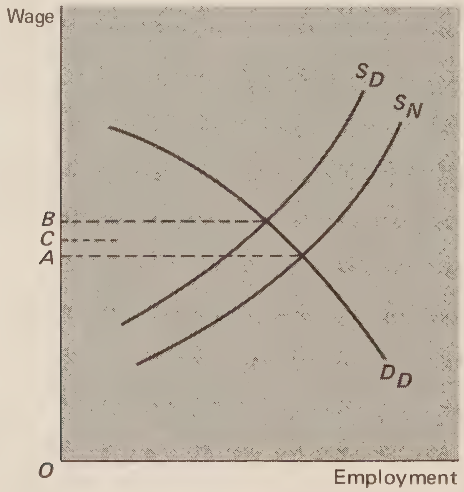


FIGURE 7–8
Rent Component of Physicians' Wages

If nurses (or others) could move freely into the medical profession, S_D would increase to the point where the wage differential between doctors and nurses is *AC*. Consequently, rent exists because of barriers to occupational mobility. These barriers may occur because of the need for specialized characteristics such as ability and motivation. If only a very few people have the requisite skills, supply will be limited. On the other hand, artificial barriers, including prejudiced attitudes against certain types of individuals, may limit entry into the medical profession.

Rents may be merely temporary or they may exist in the long run. A particular skill may be in fixed supply only temporarily. Rent accruing to such resources is called **quasi-rent.** In the 1950s, when the Soviet Union launched the first Sputnik satellite, there was a sharp increase in demand for engineers and scientists. Since these occupations require a long training period, the supply of engineers and scientists was fixed in the short run.

In Figure 7–9 the short-run supply of engineers and scientists is fixed at S_1.

[7] In general, rent is the difference between the actual payment to a factor of production and the payment that would be required to attract it into service.

Because of the long training period required, no increase in the wage could induce an increase in supply in the short run. Before the Sputnik scare the demand for scientists and engineers was D_1, and the wage OA in relation to wages in other occupations presumably reflected differences in training costs, psychic rewards, and so on. When demand increased to D_2 the wage rose to OB. The increase AB can be viewed as a pure quasi-rent, since it is a bonus over and above that required to attract S_1 scientists and engineers into the profession.

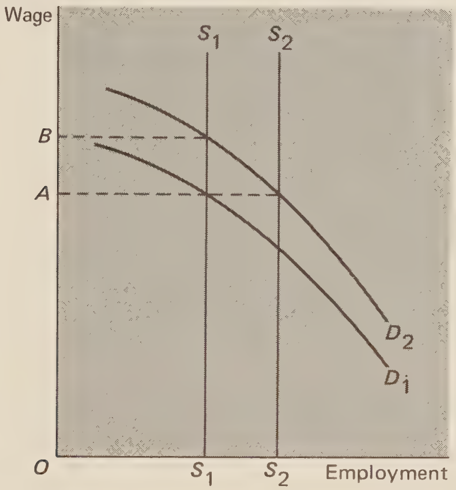


FIGURE 7–9
Increase in Demand for Engineers and Scientists

Eventually, the supply of scientists and engineers should increase in response to the high wages in that occupation. In fact, in the late 1950s and early 1960s many young people entered engineering school and pursued scientific degrees. By the mid 1960s the supply had increased to S_2, eliminating quasi-rents but maintaining a wage that reflected training costs. By 1970 the American commitment to the war in Southeast Asia and inflationary pressures resulted in a substantial cutback in the aerospace program and a concomitant decline in demand for scientists and engineers. Suppose demand had declined to D_1. If the wage were maintained at OA to reflect training costs relative to other professions, the demand would be S_1 while the supply would remain at S_2. Thus $S_2 - S_1$ scientists and engineers would be unemployed. Indeed, in the early 1970s the unemployment problem among scientists and engineers was considerable. Many of them accepted substantial cuts in pay or moved to occupations for which their skills were wasted. Unless the demand for scientists and engineers increases again, they may never recover their training costs.

Rents may be enjoyed forever if there is a permanent barrier to the acquisition

of certain attributes and if demand continues indefinitely. Location advantages and fertility of land are good examples of productive resources that can earn rent indefinitely, provided demand remains high. Property in the middle of a large city was traditionally expected to command a higher rent than land in outlying areas. In some cities, however, because of crime in the inner city and the movement to the suburbs, this trend has been reversed.

It is important to note that the existence of rent is not inconsistent with the marginal productivity principle. In fact, an increase in rent as demand increases is a signal reflecting the scarcity of that factor of production. If it is a quasi-rent, the increase will have the effect of attracting resources into that occupation, and it will eventually disappear. Thus rents are an important allocative device and must exist if optimal resource allocation is to be achieved.

Some economists have argued, however, that since rents are a payment in excess of that required to keep a resource in its current occupation, they can be taxed away without affecting the allocation of resources. This is clearly not the case for quasi-rents. There may be some merit to the proposal if the rent accrues in the long run. However, even if rents can be taxed without affecting resource allocation, this does not imply that their existence is inconsistent with Pareto efficiency or with the marginal productivity principle.

7–4 THE PRODUCT EXHAUSTION THEOREM

The marginal productivity principle of distribution states that all factors of production will be paid in proportion to their marginal product. In order for this principle to work in practice, it must be demonstrated that the output generated by the factors of production will be sufficient to remunerate them their marginal value product.

Product exhaustion occurs when the entire output is distributed to factors of production that are paid their marginal value product. There are two possible assumptions that will product the result known as product exhaustion. The first is an assumption about the production functions of individual firms; the second is an assumption about the conditions for long-run equilibrium for the industry in which the firm operates.

Consider a single firm producing an output that can be sold at a fixed (market) price and hiring factors of production at fixed prices. In this situation the conditions for product exhaustion depend on the specification of the production function. The **product exhaustion theorem**[8] states that the production function must be linear homogeneous for product exhaustion to occur, that is, that if, and only if, the production function for a firm is linear homogeneous and all factors are paid the value of their marginal product, then the total product will be completely distributed to the factors of production. Algebraically,

[8] The mathematical formulation is called Euler's Theorem.

$$pq = MVP_L L + MVP_K K$$

or

$$pq = \sum_{i=1}^{n} MVP_i F_i$$

where pq is the value of output, MVP_i is the marginal value product of the ith factor, and F_i is the amount of the ith factor employed.

Suppose the production function is not linear homogeneous. If we continue to remunerate factors according to their marginal value products, what will be the relation between the total product and the amount to be distributed?

A linear homogeneous production function exhibits constant returns to scale. For a production function that is homogeneous but exhibits nonconstant returns to scale, the optimal factor proportions are invariant to the scale of operation when factor prices are fixed, but the change in output when all inputs are increased (or decreased) proportionately is nonproportional. Algebraically, a function

$$q = \phi(F_1, F_2, \ldots, F_n)$$

is homogeneous of degree k if for any scalar t

$$t^k q = \phi(tF_1, tF_2, \ldots, tF_n)$$

That is, if all inputs are increased by some amount, t, output is increased by t taken to the kth power. Euler's Theorem states that for any such function

$$kpq = \sum_{i=1}^{n} MVP_i F_i$$

If $k = 1$, this is the product exhaustion theorem. If $k > 1$, this implies increasing returns to scale, and

$$pq < \sum_{i=1}^{n} MVP_i F_i$$

If factors are paid their marginal value product under conditions of increasing returns to scale, the factor payments would be greater than the value of the output. Consequently, factors cannot be paid their marginal value product from the internal resources of the firm.

Since the marginal productivity principle of distribution is necessary to achieve Pareto optimal allocation of resources in production (in the absence of externalities),[9] a case can be made that firms producing under conditions of increasing returns should receive an outside subsidy in order that the factors can be renumerated in proportion to their marginal productivity.

---

[9] Remember that Pareto optimality does not consider the equity of the distribution principle. The marginal productivity principle of distribution achieves efficiency in production, but it does not necessarily produce a desirable distribution of income.

Suppose, on the other hand, that $k < 1$, implying that the production function exhibits decreasing returns to scale. In this case,

$$pq > \sum_{i=1}^{n} MVP_i F_i$$

and there will be some of the product left over after distribution to the factors according to the marginal productivity principle.

If decreasing returns to scale can be attributed to a factor such as management, which is fixed in supply, the undistributed residual can be viewed as rent. Neoclassical economists like Marshall and Pigou argued that such residuals could be used to subsidize increasing-returns industries.[10] In such a case, however, it would be important to ascertain if the residual were not, in fact, a quasi-rent that serves the function of attracting scarce resources into the firm or industry. If it were a quasi-rent, imposing a tax would be undesirable, since it would interfere with the allocative mechanism of the price system.

From the viewpoint of a single firm, the assumption of a linear homogeneous production function is necessary for the practical application of the marginal productivity principle. If there is too little output to go around, as in the case of increasing returns, distribution cannot be based on marginal productivity without an outside subsidy. If there is some left over, as with a firm operating under conditions of decreasing returns, we need a theory to explain how the residual is distributed.

Another way of stating the condition for product exhaustion is that there must be no excess profits or losses when all factors are remunerated according to their marginal value products. Euler's Theorem referred to a single firm for which product and factor prices were assumed to remain fixed as output varied. Suppose, however, we consider changes in output for the industry as a whole. In the market model, firms are free to enter or leave any industry in response to expected profits or losses. Since the size of firms is presumed to be small in relation to the market, firms will enter as long as they expect to make profits in excess of variable factor payments and leave when they expect losses without recognizing their impact on market prices and costs. However, the entry of many firms will drive down market prices and drive up costs (because of the increased demand for inelastically supplied factors) until profits are reduced to zero. At this point there will be neither entry nor exit and a long-run, zero profit equilibrium will be established. Product exhaustion then occurs regardless of the specification of the production function.

Thus the assumption of linear homogeneity is not necessary to produce product exhaustion in the long run in the market model. Nevertheless, for short-run situations, the assumption of linear homogeneity is necessary. Furthermore, Euler's Theorem is useful for demonstrating the implications of the marginal productivity principle of distribution when production is taking place under con-

[10] For an evaluation of this proposal see Howard S. Ellis and William Fellner, "External Economies and Diseconomies," *American Economic Review, XXXIII* (September 1943), pp. 493–511.

ditions of increasing or decreasing returns in the short run. In such cases, the marginal productivity principle, although a necessary condition for Pareto efficient resource allocation (in the absence of externalities), breaks down as a complete theory of distribution.

7–5 FACTOR SHARES AND THE PRODUCTION FUNCTION

If we assume that resources are paid in proportion to their marginal products, we can analyze factor price differentials among occupations, industries, and different groups of workers. We may also be interested in the way in which the total product is distributed among the different factors of production, and in how these **functional income shares,** that is, shares of the product going to the different productive factors, change over time and in response to changes in factor prices.

President Nixon applied the concept of wage-price guidelines as an anti-inflation measure in the early 1970s. The idea was to prevent wages from rising faster than marginal productivity in order to prevent prices from rising at all. If labor is the only variable factor of production and if firms are maximizing profit, then the condition

$$p = \frac{w}{MP_L}$$

is attained.[11] If wages rise no faster than marginal productivity, prices should not increase.

The guidelines policy was applied only to wages, and the cost of other productive factors was not controlled. If other factor prices increase, there should be some incentive to substitute labor for other factors if labor is not the only variable factor of production. The longer a guidelines policy is in effect the more likely such substitution will occur.

Labor unions argued that the guidelines policy favored capital, since wages were the only factor price that was controlled. Firms argued that the guidelines harmed profits, or capital's share of the product, because prices were not allowed to increase as fast as wages. Can we use the theory of production to analyze the effect on relative factor shares of moving to a new expansion path in response to changing factor price ratios?

The share of the product going to a factor is the factor price multiplied by the amount of the factor employed. Thus the share of labor relative to the share of capital, $R_{L/K}$, is

$$R_{L/K} = \frac{p_L L}{p_K K}$$

If we assume that factors are paid their marginal product, so that

[11] This assumes the competitive market model. We will drop that assumption in chapter 9.

$$MRS_{KL} = \frac{p_L}{p_K}$$

then the effect on relative factor shares of a change in factor prices depends on the **elasticity of substitution** of the factors in the production function. The elasticity of substitution is the per cent change in the capital-labor ratio relative to a per cent change in the marginal rate of substitution, that is,

$$\sigma = \frac{\Delta(K/L)}{K/L} \div \frac{\Delta MRS_{KL}}{MRS_{KL}} = \frac{MRS_{KL}}{K/L} \times \frac{\Delta(K/L)}{\Delta MRS_{KL}}$$

where σ is the elasticity of substitution.

If factors are paid in proportion to their marginal products, a per cent change in the capital-labor ratio divided by a per cent change in relative factor prices is the elasticity of substitution, or

$$\frac{\Delta(K/L)}{K/L} \div \frac{\Delta(p_L/p_K)}{p_L/p_K} = \frac{\Delta(K/L)}{K/L} \div \frac{\Delta MRS_{KL}}{MRS_{KL}} = \sigma$$

If $\sigma = 1$, and if factors are paid their marginal product,

$$\frac{\Delta(K/L)}{K/L} \div \frac{\Delta(p_L/p_K)}{p_L/p_K} = 1$$

or

$$\frac{\Delta(K/L)}{K/L} = \frac{\Delta(p_L/p_K)}{p_L/p_K}$$

A per cent change in the capital-labor ratio is exactly offset by an equal per cent change in factor prices, so there is no impact on relative factor shares. Suppose, for instance, wage-price guidelines reduce wages relative to the cost of capital by 5 per cent. This will encourage a substitution of labor for capital, reducing the capital-labor ratio by 5 per cent. Since relative factor shares can be expressed as

$$R_{L/K} = \frac{p_L L}{p_K K} = \frac{p_L}{p_K} \times \frac{L}{K}$$

it is clear that an equal decrease in p_L/p_K and K/L will have no impact on $R_{L/K}$. Thus a guideline policy should leave relative shares unaffected if the underlying aggregate production function exhibits unitary elasticity of substitution and if the factors are being remunerated in proportion to their marginal products.

Empirical studies of the United States over long periods of time have shown very little change in relative factor shares despite rather wide fluctuations in relative factor prices and in the relative factor-intensiveness of aggregate production.[12] This has led many economists to believe that the aggregate production

[12] See Robert M. Solow, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics, XXXIX* (August 1957), pp. 312–320.

function must have the property of unitary elasticity of substitution. The Cobb-Douglas production function

$$X = AL^{\alpha}K^{1-\alpha}$$

has this property. No matter which expansion path is chosen, if L and K are paid in proportion to their marginal products, relative factor shares are always the same.[13]

Other studies have shown that since 1950, as production techniques have become more capital-intensive, the share of national income going to labor has increased slightly.[14] This implies that the elasticity of substitution must be less than 1, since the per cent increase in p_L relative to p_K must have been greater than the per cent decrease in L/K. On the basis of these studies economists have developed a production function of the form

$$X^{-\beta} = aL^{-\beta} + bK^{-\beta}$$

where

$$\sigma = \frac{1}{1+\beta} \qquad \text{and} \qquad \frac{p_L}{p_K} = \frac{b}{a}\left(\frac{L}{K}\right)^{1/\sigma}$$

This function has a constant elasticity of substitution equal to $1/(1 + \beta)$ and is linear homogeneous.[15]

Let us return to the question of the impact of wage-price guidelines on the American economy. If the elasticity of substitution of the aggregate production function is unitary, then the guidelines should have no impact on relative factor shares. Suppose, on the other hand, that σ is somewhat less than 1. As capital costs rise relative to wages, labor is substituted for capital. But the rise in p_K/p_L will not be completely offset by an increase in L/K. Consequently, the impact of the guidelines would be to increase the profit share, that is, the share going to capital.

Observers of the guidelines policy of the Kennedy-Johnson administrations for the period 1962–66 have estimated that the guidelines favored profits relative to wages. The rate of increase of unit labor costs, w/AP_L, declined by 2 per cent, while the rate of increase of prices declined by only 1.4 per cent.[16] Although these results can be attributed to other factors such as the use of pricing policies other than profit maximization, they are also consistent with the evidence that the elasticity of substitution in aggregate production for the United States economy is less than 1.

---

[13] The actual share going to labor is α and the share going to capital is $1 - \alpha$. Since $\alpha + (1 - \alpha) = 1$, the total product is exhausted in factor payments. This is because the Cobb-Douglas function is linear homogeneous.

[14] See, for example, K. J. Arrow et al., "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics, XLIII* (August 1961), pp. 225–250.

[15] If $\beta = -1$, the function is undefined and the Cobb-Douglas function is used.

[16] Gary Fromm, "The Wage-Price Issue: The Need for Guideposts," in *Joint Economic Committee Hearings* (Washington, D.C.: Government Printing Office, 1968), pp. 3–7.

The Impact of Technological Change

If we are interested in explaining how relative factor shares change over time, we must take the effect of technological change into account. Neutral technological change leaves the relation between *MRS* and *K/L* unaffected, so that it will not affect relative factor shares.

If technological change is labor-using, MRS_{LK} will fall, implying that MP_L rises relative to MP_K for a given expansion path. If factors are paid their marginal product, the share going to labor will rise when labor-using technological change takes place. One explanation for the increase in labor's share in recent years is that technological change has been of the labor-using variety. Empirical evidence also supports this view.[17]

7–6 EVALUATION OF THE MARGINAL PRODUCTIVITY PRINCIPLE OF DISTRIBUTION

In recent years there has been a growing dissatisfaction with the marginal productivity approach to the theory of income distribution. The critics fall into two groups. Some attack the theory on the grounds that marginal productivity cannot be measured by the firm and that employers use other, more subjective, criteria for determining how factors are remunerated. Others do not deny the validity of the marginal productivity principle but argue that this analysis is not the most useful way to analyze the problem of income inequality.

We have seen that it is not necessary to be able to determine the marginal product in a pure sense (the output attributable to the last unit of input), provided that values can be imputed to the productive factors in the form of shadow prices. However, as the economy becomes increasingly service-oriented and less goods-oriented, the measurement of output becomes more difficult and leaves itself open to subjective evaluation on the part of the employer. In 1971, 85 per cent of the female labor force and 54 per cent of the male labor force held white-collar jobs or were in service occupations.

If some workers are viewed as "inferior" to others, their employers may perceive their marginal product to be lower than that of "superior" workers. Once marginal productivity is defined in subjective terms, the possibility for discriminating on the basis of the personal characteristics of the worker is increased.

Other critics of the marginal productivity approach say that the important issue is what determines real (or perceived) productivity initially. They argue that labor markets are "stratified" on the basis of age, race, and sex because of the inability of some groups to acquire the necessary skills and appropriate cultural accoutrements for high-paying jobs, and that remuneration may be based on unjustified social attitudes towards blacks, women, and young people that prevent them from competing effectively with more privileged groups.

---

[17] See C. E. Ferguson, "Substitution, Technical Progress, and Returns to Scale," *American Economic Review, Papers and Proceedings, LV* (May 1965), pp. 296–305.

Normative Aspects of the Theory of Distribution

The marginal productivity principle describes how income is distributed, not how it should be distributed. Thus it is not intended to be a normative theory. Although some neoclassical economists attempted to make a case for the desirability of a distributive mechanism that rewards in proportion to one's contribution to production, it is clear that differences in the initial distribution of resources and differences in the ability and opportunity to acquire skills render such an interpretation subject to question. Marginal productivity also depends on the use of other, cooperative factors of production such as capital and natural resources, from which some individuals are barred access, for geographic or other reasons. Furthermore, the marginal productivity principle as a normative theory of distribution is not necessarily consistent with the principle of maximizing social welfare. It would imply that the most productive people should get more because they receive more psychic gratification from consumption.

The marginal productivity principle of factor remuneration is necessary for the attainment of Pareto efficiency in production. Only if factors are paid their marginal value product will firms select the least cost expansion path (production technique) and select that scale of operation that equalizes the marginal rate of transformation for all firms in the society. This ensures that all goods are produced at their minimum opportunity cost in relation to all other goods. The product exhaustion theorem establishes the feasibility of the marginal productivity principle under nonincreasing returns to scale.

But the marginal productivity theory does not solve the problem of determining the optimal distribution of income once Pareto efficiency is established. Despite a considerable academic literature on this issue, most economists agree that it remains a normative question, requiring a subjective value judgment. We will return to the income distribution controversy in chapter 11.

QUESTIONS FOR STUDY AND REVIEW

1. Economic theory tells us that wages are determined by the marginal productivity of labor. How can this be reconciled with the fact that wage rates of several kinds of workers have increased over the decades while their output per man-hour has hardly changed at all? Wages of house painters, who use the same techniques they used half a century ago, have increased approximately as much as those of workers whose output per man-hour has doubled several times. Explain the apparent paradox.
2. Labor force participation rates for married women over 25 have increased dramatically since 1960. Using marginal productivity theory, analyze the potential impact of this trend on male-female wage differentials.
3. Some economists have argued that income distribution in the United States is determined by factors other than marginal productivity. What are the implications of this viewpoint for economic efficiency in the producing sector?
4. A university library has two parking lots. Lot A is adjacent to the library, while lot B is a five-minute walk away. Lot A is always filled, while lot B usually has several unused

spaces. The cost of maintaining the two lots is negligible. The university is considering charging a 50-cent parking fee for use of lot A. Students protest, arguing that since the university incurs no cost for maintaining the parking lots, this is just a way of extracting more money from students. Do you agree? What would be the impact of such a parking fee? Would the fee increase or reduce the efficiency of resource allocation?

5. "The elasticity of the market demand curve for a factor of production depends on the elasticity of demand for the product." Explain why this is so. Is this inconsistent with the marginal productivity principle?
6. A recent study has shown that the elasticity of factor substitution in many countries of Latin America is about 0.75. If this is the case, what trends do you project for the functional distribution of income during the process of industrialization?
7. How can technological change affect the demand for a factor of production? How can it cause unemployment? What market forces tend to eliminate such unemployment? What factors may impede the operation of those forces?

ADDITIONAL READING

Becker, Gary. *Human Capital*. New York: Columbia University Press, 1964.

Buchanan, Daniel H. "The Historical Approach to Rent and Price Theory," *Economica, IX* (June 1929), pp. 123–155.

Hicks, John R. *The Theory of Wages,* 2d ed. New York: St. Martin's Press, 1963, chapters 1–7.

Kaldor, Nicholas. "Alternative Theories of Distribution," *Review of Economic Studies, XXIII* (1955–56), pp. 83–100.

Machlup, Fritz. "On the Meaning of the Marginal Product," in *Explorations in Economics* (1936), pp. 250–263. Reprinted in the American Economic Association. *Readings in the Theory of Income Distribution*. Philadelphia: Blakiston, 1951, pp. 158–174.

Marshall, Alfred. *Principles of Economics*. London: Macmillan and Co., Ltd., 1959, book V, chapters 6, 8.

Robertson, Dennis H. "Wages-Grumbles," in *Economic Fragments* (1931), pp. 42–57. Reprinted in the American Economic Association. *Readings in the Theory of Income Distribution*. Philadelphia: Blakiston, 1951, pp. 221–236.

Solow, Robert M. "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics, XXXIX* (August 1957), pp. 312–320.

Thurow, Lester. *Poverty and Discrimination*. Washington: The Brookings Institution, 1969.

8 Market Equilibrium

The theory of the household provided an explanation of the determinants of product demand and factor supply. The theory of production can be used to analyze the supply of products and the demand for factors of production. These models have also demonstrated the necessary preconditions for the existence of efficiency in consumption and production.

In chapters 2 and 3 we examined the necessary conditions for global efficiency. First, the marginal rate of substitution between two goods in consumption must be equal to the opportunity cost of one in terms of the other, the marginal rate of transformation. Assuming consumers and producers both view prices as fixed, consumers will purchase goods until the marginal rate of substitution is equal to the price ratio, that is,

$$MRS_{XY} = \frac{p_Y}{p_X}$$

and firms will produce goods until the marginal rate of transformation is equal to the price ratio, that is,

$$MRT_{XY} = \frac{p_Y}{p_X}.$$

Thus, when goods are traded at fixed prices, the efficiency condition

$$MRS_{XY} = \frac{p_Y}{p_X} = MRT_{XY}$$

is achieved.

Not only must the marginal condition be met, however, but there must be no excess supply of, or demand for, any good or service if efficiency is to be achieved. If goods are traded in markets, efficiency requires that the markets be in equilibrium, that is, that the quantity demanded is equal to quantity supplied at the market price.

In factor markets the marginal condition for efficiency states that the marginal dollar product of a factor service to the firm must be equal to the marginal rate of substitution of the factor for income for

the household. If a fixed factor price is established, the marginal condition will be met, since

$$MRS_{f\$} = p_f = MP_{f\$}$$

In addition to the marginal condition there must be no excess supply of, or demand for, the factors. If factors are traded in markets, the markets must be cleared to achieve efficiency.

Not only is market equilibrium a precondition for efficient resource allocation but equilibrium analysis is an important methodological device in economic theory. The assumption that markets are always in equilibrium or are approaching equilibrium fairly quickly greatly facilitates theoretical analysis but often produces misleading or incorrect results.

In this chapter we examine the way in which markets establish equilibrium. This will provide a basis for evaluating both the efficiency of resource allocation in a market system and the usefulness of the equilibrium concept in economic theory.

8–1 THE CONCEPT OF MARKET EQUILIBRIUM

A market is said to be in equilibrium when the quantity supplied equals the quantity demanded. The price that clears the market is the equilibrium price. If there is a single equilibrium price, no other price will clear the market. If the prevailing price is above equilibrium, as p_1 in Figure 8–1, the quantity supplied, q_{S1} will exceed the quantity demanded, q_{D1}. Since firms cannot dispose of all their output at p_1, selling prices will be adjusted downward until the market is cleared

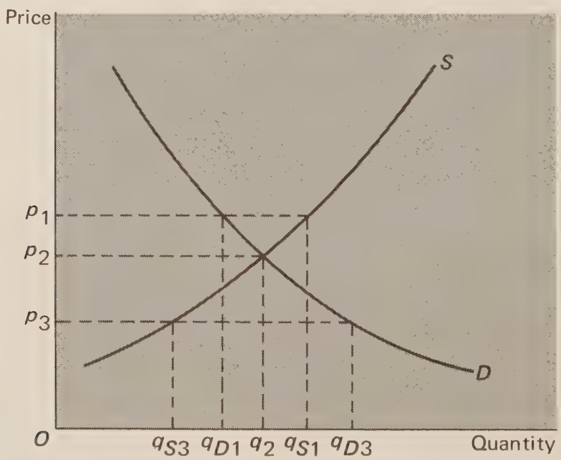


FIGURE 8–1
Market Equilibrium

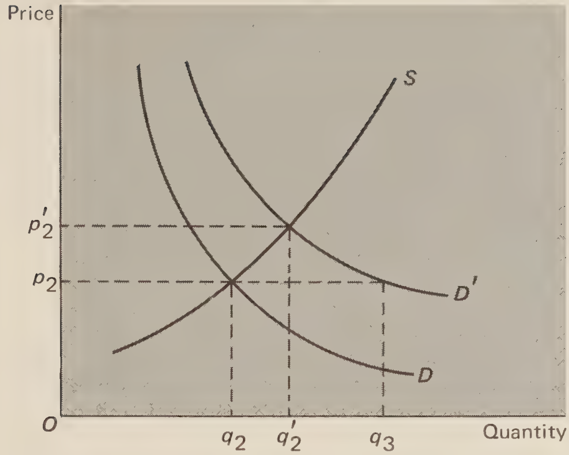


FIGURE 8-2
Impact of a Shift in the Demand Curve on Market Equilibrium

at q_2, and p_2, the equilibrium price, is established. If the price is p_3, below equilibrium, quantity demanded, q_{D3} will exceed quantity supplied, q_{S3}. Buyers will not be able to purchase all they want at p_3 and will bid prices up to p_2, which will clear the market. When p_2 is reached, the amount buyers wish to produce will exactly equal the amount firms wish to sell, so there will be no effort on the part of buyers or sellers to readjust prices. Consumers and producers can buy and sell all they wish at the market price, and that price will remain fixed as long as there are no changes in demand or supply.

Once the equilibrium price is established, there will be no readjustment of price unless there is a change in the underlying determinants of demand or supply. An increase in income, for instance, will cause the demand function to shift to D', as shown in Figure 8-2. The new equilibrium price is p_2'. The former equilibrium price, p_2, now is associated with excess demand and so will not be maintained. Buyers will bid up prices until p_2' is achieved. A general definition of market equilibrium is the attainment of a position from which there is no tendency to move, subject to the underlying determinants of the supply and demand functions.

The attainment of market equilibrium is essential to the achievement of Pareto efficiency, since it is the equilibrium price that equates the marginal rate of substitution and marginal rate of transformation in the market model. In the real world, however, the underlying determinants of market demand and supply change frequently, and consequently it is important to investigate the speed with which equilibrium is reestablished. Equilibrium analysis must include a dynamic analysis of market behavior in disequilibrium situations. Finally, the theory of general equilibrium examines the question of whether equilibrium is likely to be established in all interrelated product and factor markets simultaneously.

8–2 COMPARATIVE STATICS AND THE CORRESPONDENCE PRINCIPLE

Not only is the attainment of equilibrium necessary to achieve efficient resource allocation but the equilibrium concept is central to the methodology of economic theory. Suppose an increase in import quotas causes an increase in the supply of meat, as shown in Figure 8–3. When the supply schedule shifts from S to S', the equilibrium price falls from p_2 to p_1. This implies that if the market were initially in equilibrium at p_2 and if equilibrium were restored, the effect of the increase in supply would be a decline in the market price.

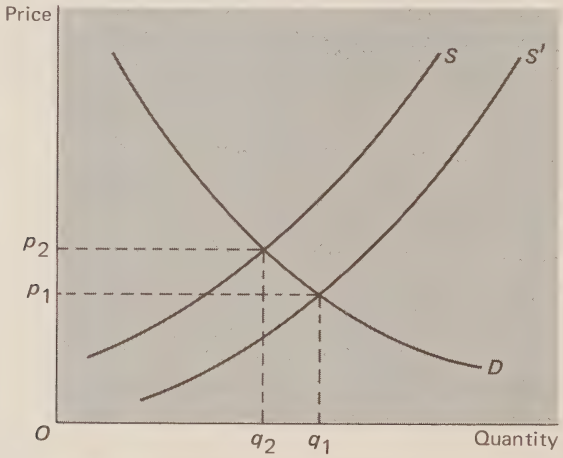


FIGURE 8–3
Impact of a Shift in the Supply Curve on Market Equilibrium

The examination of equilibrium positions associated with different supply and demand functions is called **comparative statics.** In many instances, a comparative statics methodology is a good way to examine market behavior. If we want to predict the impact of an increase in meat quotas on market price, comparative statics would indicate that market price would fall if the market supply and demand curves had the normal shape. The approach assumes, however, that the new equilibrium will actually be achieved.

Suppose consumers are slow to adjust their purchases to changes in meat prices. When supply increases from S to S' and there is no initial increase in demand, price will fall to p_2, as shown in Figure 8–4. When price is p_2, quantity demanded will eventually increase to q_2, causing price to rise to p_3, which brings forth an increase in the quantity supplied. Demand will fall to q_3, causing price to fall below p_2 and causing a reduction in quantity supplied. The new equilibrium price, p_1', is never attained, and market price will continue to fluctuate in cycles of increasing amplitude. Excess demand and supply increase in magnitude, resulting in increasing severity of shortages alternating with surpluses.

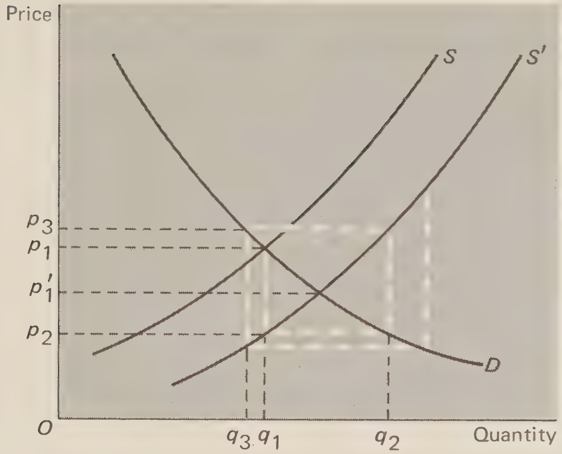


FIGURE 8-4
Dynamic Analysis of an Increase in Supply (Quantity Lag)

Comparative statics, which implies that meat prices will decline in response to an increase in supply, is not a useful or correct approach for studying price behavior in this particular market.

On the other hand, price may be slow to adjust to the initial increase in supply. In that case, quantity supplied increases from q_1 to q_2, as shown in Figure 8-5. At q_2 there is an excess supply, and eventually producers will lower prices to p_2 to eliminate unwanted inventories. Supply will fall to q_3, raising price to p_3. Price fluctuations become smaller and smaller until eventually the new equilibrium price, p_1' is established.

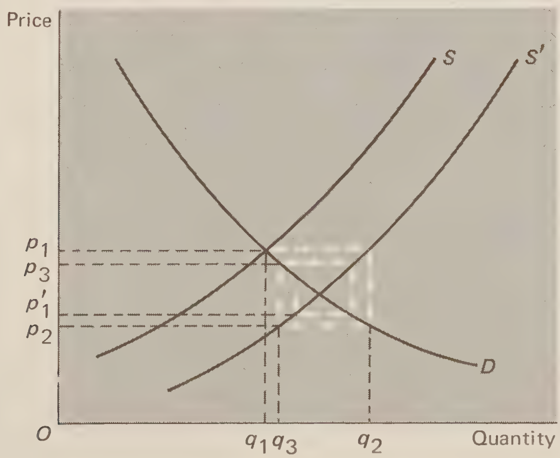


FIGURE 8-5
Dynamic Analysis of an Increase in Supply (Price Lag)

In the case of a price lag, the comparative statics approach would have been fruitful. A quantity lag would produce a situation in which equilibrium is never attained.

Although the comparative statics approach is often used for analyzing the impact of a change in demand and supply on market price and output, it should never be used without first examining the process by which the market adjusts to a disequilibrium situation. Samuelson has called this the **correspondence principle.**[1] Corresponding to every problem in comparative statics is an underlying dynamic adjustment process that must be investigated before the analysis of equilibrium will yield fruitful results.

It is important to recognize that the correspondence principle is not inconsistent with the essentially static nature of equilibrium. If an equilibrium price is achieved, there will be no tendency for readjustment unless there is a change in the underlying demand and supply functions. In Figure 8–4, if p_1' were ever achieved, the market would be cleared and there would be no readjustment of price. The correspondence principle, however, focuses on the question of whether p_1' will ever be achieved.

In Figure 8–4 the equilibrium price, p_1', and the corresponding output, q_1', are said to be **unstable.** They are unstable because the actual market price and output traded never reaches this point. In Figure 8–5 the equilibrium (p_1', q_1') is **stable** because the market adjusts to it. Notice that the stability or instability of equilibrium is not a characteristic of the equilibrium itself but of the underlying dynamic process. Clearly, stability analysis is an important part of comparative statics, or equilibrium analysis. While comparative statics involves examining equilibrium under a variety of market conditions, stability analysis is concerned with the dynamic process of adjustment to disequilibrium situations. Once market stability is ascertained, comparative statics becomes a reasonable approximation to actual market behavior.[2]

8–3 EQUILIBRIUM IN PRODUCT AND FACTOR MARKETS

Deferring the question of stability for a time, we will focus on the comparative statics approach. We assume that all equilibria are stable and that the process of adjustment is fairly rapid.

---

[1] Paul A. Samuelson, *Foundations of Economic Analysis* (Cambridge, Mass.: Harvard University Press, 1947), p. 5.

[2] Comparative statics also implies the adjustment process is not excessively slow. In the Keynesian model downward rigidity of wages and interest rates impedes the process of adjustment to an excess supply of labor within a reasonable time frame, even if the national income equilibrium is stable. For a discussion of stability conditions for macroeconomic equilibrium see Nancy S. Barrett, *The Theory of Macroeconomic Policy* (Englewood Cliffs, N.J.: Prentice-Hall, 1972), pp. 288–292.

Economic Activities of Labor Unions

Consider the way in which a labor union can influence market conditions. Figure 8–6 shows the market demand and supply curves for labor in the steel industry. The market demand curve is the horizontal sum of the individual marginal value product curves of the steel companies, taking into account the impact of changing levels of market demand and output on market price (see chapter 7).[3] Given prevailing demand conditions, the firms will be induced to hire more workers only if the wage is reduced. This is because the marginal productivity of labor will fall as more workers are hired and because product prices will fall as more output is produced.

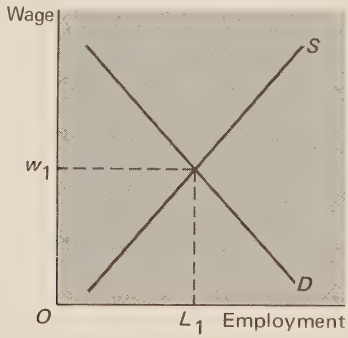


FIGURE 8–6
Labor Market Equilibrium

The labor supply curve reflects differences in psychic rewards and training costs associated with the occupation as opposed to other occupations as well as the marginal rate of substitution of income for leisure. An increase in wages will attract more workers into the industry, since the occupation will become more attractive relative to other occupations.[4] The equilibrium level of employment is L_1 and the equilibrium wage is w_1.

Suppose the steelworkers' union is dissatisfied with this particular equilibrium position and attempts to change conditions in the labor market. The union could have a number of objectives, some of them mutually exclusive.[5] These include:

[3] We will assume, for simplicity, that individual steel firms are not large enough to affect market price. We will drop this assumption in chapter 9.

[4] An increase in wages for the economy as a whole might cause a decline in labor force participation if leisure is an inferior good or, alternatively, if the marginal utility of income is declining very rapidly. This is unlikely to occur in an individual occupation, however, since an increase in wages will attract labor from other occupations.

[5] For a discussion of the economic and noneconomic objectives of labor unions see Allan M. Cartter, *Theory of Wages and Employment* (Homewood, Ill.: Richard D. Irwin, 1959), chapter 7.

1. Increasing wages for union members.
2. Increasing wages for union members with a minimum of unemployment.
3. Maximizing total wages for union members.

Suppose the union wishes to raise wages from w_1 to w_2, as shown in Figure 8–7. One way of achieving this goal is to obtain the increase through collective bargaining. If wages rise from w_1 to w_2, demand will decline from L_1 to L_0 as shown in part A of Figure 8–7. Supply will increase from L_1 to L_2. Thus there will be unemployment of $L_2 - L_0$ workers.

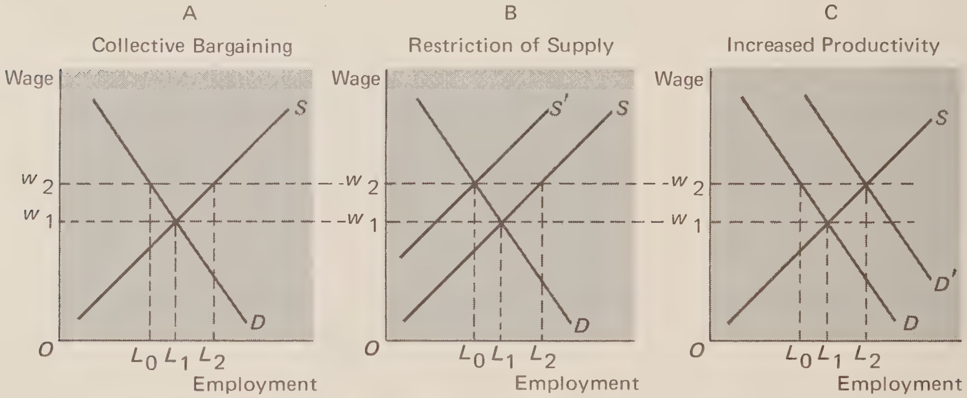


FIGURE 8–7
Methods of Obtaining a Wage Increase

A second approach often used by labor unions is to restrict supply to the industry. In part B of Figure 8–7, if the supply function could be shifted to S', then w_2 would be the new equilibrium wage. In the early days of trade unionism, unions won closed shop agreements, which barred from employment persons who were not members of the union. Thus unions could control the supply of labor by limiting union membership. Closed shop agreements are now illegal. However, unions can·still restrict labor supply by requiring specialized training and apprenticeships. We have seen how training costs increase the supply price of labor, causing the supply curve to shift leftward.

Notice that in both cases A and B, when the wage rises to w_2 only L_0 labor is employed. In case A this leaves $L_2 - L_0$ workers unemployed. In case B the supply of labor falls by $L_2 - L_0$. Case A is a disequilibrium situation, however. If employment remains a problem in the industry, unemployed workers may look for jobs elsewhere, thus causing the supply curve to shift to the left. If the labor in a particular industry is mobile, the union will not have to resort to restrictive practices to limit the labor supply. The labor supply will decline automatically when the union uses collective bargaining to push wages above equilibrium levels. Samuelson notes that this analysis is consistent with the observation that unions

have been liberalizing restrictions on union membership at the same time as wage rates were being pushed up by collective bargaining.[6] It is not necessary to enforce restrictions on labor supply if high wages can be maintained by collective bargaining.

Although the unemployment caused by maintaining wage levels above equilibrium in a particular occupation or industry may be only temporary or **frictional,** the same cannot be said when such a practice is economywide. Full-employment is restored in the steel industry by steelworkers moving to other industries, causing a leftward shift in the supply curve of labor to the steel industry. For the whole economy, if the supply of labor exceeds the demand at the prevailing wage, the economy can remain in a permanent state of unemployment. The unemployment can be eliminated only by a reduction in wages or by a shift in the demand or supply curves for labor for the economy as a whole.[7]

If wages are to rise throughout the economy without increasing unemployment, the demand for labor must increase. Keynes argued that increasing aggregate demand is a more effective policy for eliminating unemployment than wage cuts. Unless demand increases, full employment equilibrium can only be established at a lower wage, reflecting a lower marginal value product of labor. This means that real income of the labor force will be lower than if demand had increased and caused a rise in the marginal value product and the real income of the labor force.

If unions can engineer an increase in demand for labor, as shown in part C of Figure 8–7, an increase in wages can be accomplished without unemployment, even in the short run. If demand increases from D to D', wages will rise to w_2 and employment will increase to L_2. If product demand is fairly elastic, measures to increase labor productivity will increase the demand for labor. In such industries unions may take an active role in encouraging training and educational programs and working conditions conducive to increased productivity. Unions may even support some types of technological change. If product demand is inelastic, improved productivity might actually reduce the demand for labor. In such industries unions may turn their attention to increasing product demand. Some unions have gone so far as to supplement the advertising efforts of their employers. The "union label," which advertises the product as being manufactured by union workers, is a notable attempt to increase product sales among union members throughout the economy.

Suppose unions are interested in maximizing total wages for union members. In this case, the union would attempt to set the wage in the unit-elastic portion of the demand curve, at w_2 in Figure 8–8. Above w_2 wage reductions would increase employment by a greater proportion, resulting in increased total wages. Below w_2 wage increases would reduce employment by a smaller proportion,

[6] Paul A. Samuelson, *Economics,* 8th ed. (New York: McGraw-Hill, 1970), p. 561.

[7] Keynes noted that a decline in real wages is required to reestablish equilibrium. Such a decrease is not always achieved by a reduction in money wages. For a further discussion see Barrett, *The Theory of Macroeconomic Policy,* chapter 9.

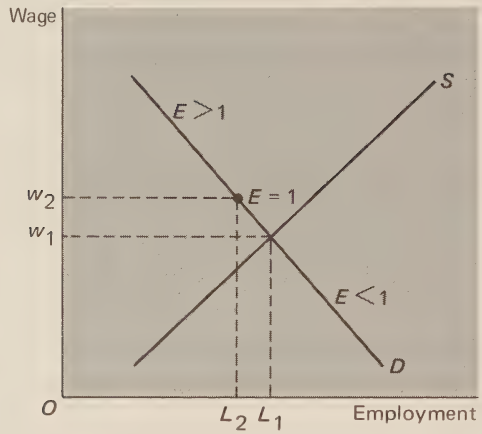


FIGURE 8–8
Maximizing Total Wages for Union Members

again resulting in increased total wages. The wage w_2 could be attained by collective bargaining or by restrictions in labor supply.

Another way to increase the total wage bill is to cause the demand curve for labor to shift rightward. Setting wages in the unit-elastic portion of the demand curve will maximize wages only if the demand curve is fixed. Rightward shifts in the demand curve will always cause an increase in total wages. However, for any demand curve, the unit-elastic range represents the maximum total wage bill.

Equilibrium in Product Markets

The market demand curve for a product is the horizontal sum of the demand curves of individual households. If the good is not an inferior good for a majority of households (weighted by their expenditures on the good), the demand curve will be negatively sloped. This implies that price reductions are required to induce an increase in demand.

The market supply curve is the horizontal sum of the marginal cost curves of the firms producing the good. Marginal costs rise more rapidly in the short run than in the long run. Consequently, the short-run supply curve is steeper than the long-run supply curve. Since marginal costs increase with output when firms are maximizing profits, the short-run supply curve is positively sloped.

In the long run the supply curve is likely to be much flatter. If firms in the industry can expand at constant costs in the long run, then the slope of the long-run supply curve may depend on interactions among the cost functions of the firms as the industry expands. If there are external economies in the industry, prices may actually fall as output increases over a period of time. Such externalities could be *pecuniary* or *technological*. An externality is present if the expansion of one firm's output affects production costs for other firms in the industry. A

pecuniary externality occurs if the cost of factor inputs changes as the industry expands. Technological externalities occur when real costs of production change because of such factors as better organization of factor markets and diffusion of improved technological knowledge.

Consider the early expansion of the automobile industry. Pecuinary externalities occur if the cost of factor inputs declines as the industry expands. Economies of scale (internal economies) in the steel and rubber industries resulted in falling materials costs once automobiles were produced in sufficient numbers to warrant large-scale production techniques for steel and rubber. Diffusion of labor

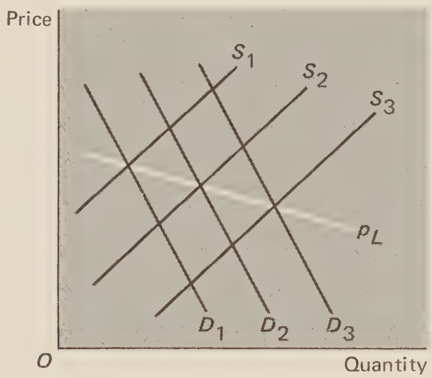


FIGURE 8–9
Effect of Shifts in Demand and Supply on Market Price

skills in the automobile industry as well as diffusion of technology produced technological externalities. These caused the marginal cost curves of the individual firms to shift to the right and the supply price for a given level of output to fall.

If the short-run supply curve shifts rightward faster than demand increases, market price will fall as the industry grows. This is seen in Figure 8–9. As demand increases from D_1 to D_2 to D_3, the supply curve is also shifting rightward, from S_1 to S_2 to S_3. The locus of short-run market equilibria, p_L, traces the pattern of long-run price behavior in that industry. If demand increases faster than supply, prices will rise in the long run as the industry expands.[8]

The impact of falling prices may be harmful or beneficial, depending on the elasticity of demand for the product. The elasticity of demand for automobiles was quite high, so that falling prices were more than offset by increases in quantity demanded. Sales and profits in the automobile industry increased as the industry grew.

[8] For a discussion of the relevance of shifting demand and supply to empirical studies of market behavior see E. J. Working, "What Do Statistical Demand Curves Show?" *Quarterly Journal of Economics, XLI* (February 1927), pp. 212–235.

On the other hand, rapid technological change in agriculture also resulted in falling supply prices for many food products. Since the demand for food is inelastic, the decline in prices was not matched by an increase in the quantity demanded. Total sales declined, as did farm incomes.

Programs to Support Agricultural Products

Government programs have been established to support farm prices and increase farm incomes. There are a number of ways in which such support programs could be designed. Such programs are comparable to the techniques unions use

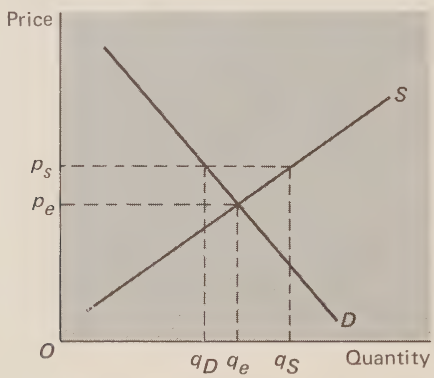


FIGURE 8–10
Impact of Agricultural Price Supports

to obtain wage increases. In Figure 8–10 the equilibrium price for agricultural goods is p_e. Suppose the government establishes p_s as the desired support price.[9] One way to raise price to p_s is for the government to offer p_s to all sellers, that is, to buy all the agricultural goods offered for sale at p_s. Consumers must compete with the government and consequently must also pay p_s. When price is p_s consumers will purchase q_D. One undesirable result of such a policy is that it encourages farmers to produce more than they would have before the support program, despite the fact that quantity demanded has actually decreased. Production has increased by $q_S - q_e$, while demand has fallen by $q_e - q_D$. The amount $q_S - q_D$ is a surplus, which the government has purchased. This surplus must be stored unless it can be distributed in domestic relief programs or sent abroad.

Most government support programs are accompanied by crop restrictions that prevent farmers from increasing acreage in response to increasing price supports. Supply could conceivably be cut back to such an extent that the equilibrium price would rise to p_s. If supply were restricted to S_R, as shown in Figure

[9] The desired support price is determined by a "parity" formula, which attempts to peg agricultural prices to other prices at a pre-World War I level.

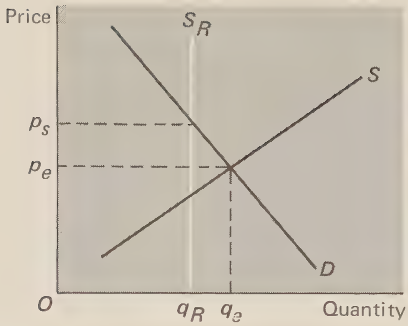


FIGURE 8–11
Crop Restriction to Raise Farm Prices

8–11, the equilibrium price of p_s would be established with no surplus. The "soil bank" programs of the 1930s and subsidies to farmers for restricting acreage in the 1950s were attempts to raise equilibrium price to desired levels. However, such radical restriction of supply was difficult to administer, and there were accusations that administrative loopholes favored large farmers at the expense of small farmers. Furthermore, there was concern that such restrictions might result in undesirable shortages in the future, particularly in years with poor climatic conditions.[10] Consequently, crop restriction is now limited to preventing further increases in acreage. This means that other programs are necessary to raise farm prices to desired levels.

Two alternative farm price support plans are shown in Figure 8–12. Say that the government purchases farm products at p_s and restricts increases in acreage in response to higher support prices or lower production costs. The supply curve, S_R becomes vertical at the existing acreage. Consumers will purchase q_D, leaving a surplus of $q_S - q_D$ when supply is restricted to S_R.

Suppose, on the other hand, that the government allows food prices to fall to their equilibrium level, p_e, and subsidizes farmers by an amount $p_s - p_e$. This plan has the advantage of lowering food prices to consumers. Because the demand for food is inelastic, consumers will actually be able to purchase more food with a smaller total outlay. Furthermore, there is a gain in consumer surplus represented by the area ABC in part A of Figure 8–12.[11]

Notice that both plans guarantee farmers the same income, Op_sDq_S. Plan B is less costly to consumers. In terms of direct purchase cost, plan A is less costly to the government, assuming that demand is inelastic. However, plan A entails the cost of storage and distribution of the surplus produce, so that the total cost of the government purchase scheme may actually exceed that of the subsidy arrangement.

[10] When acreage is fixed, supply can fluctuate dramatically from year to year in response to changing environmental conditions.

[11] See pages 89–91 for a discussion of consumer surplus.

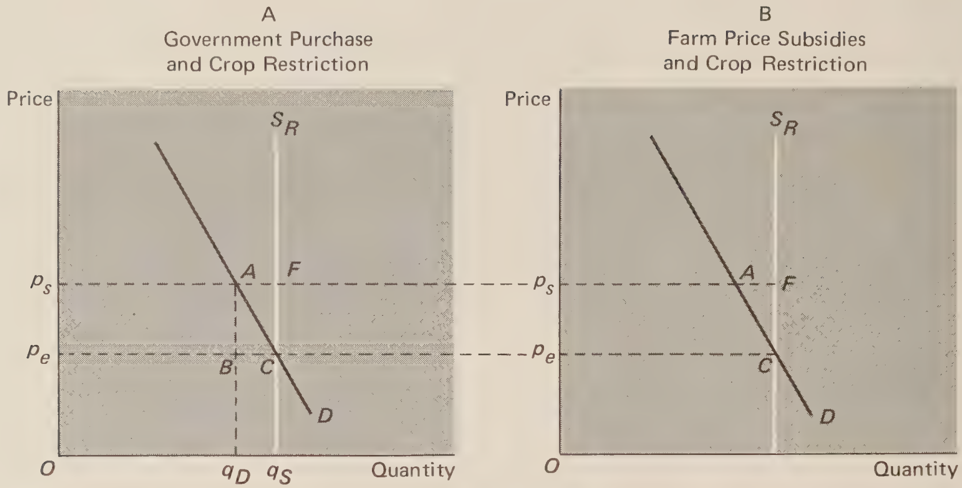


FIGURE 8–12
Alternative Programs to Support Farm Prices

8–4 STABILITY ANALYSIS

Equilibrium analysis, or comparative statics, can be used to examine many important policy issues. However, we have seen that the approach yields fruitful results only to the extent that market prices actually tend to move to their equilibrium values. If a market is in disequilibrium because of a shift in the demand or supply functions and if it tends to adjust to a new equilibrium, that equilibrium is said to be stable. The stability of an equilibrium is related to the way the market participants adjust to disequilibrium situations as well as to the nature of market demand and supply. Investigation of these market characteristics is called **stability analysis.**

Classical Stability Analysis: Instantaneous Adjustments

Consider a market in disequilibrium, that is, at the prevailing price there is excess demand or supply. We can define $E(p)$, excess demand at price p, as the difference between the quantity demanded at that price, $D(p)$, and the quantity supplied, $S(p)$, that is,

$$E(p) = D(p) - S(p)$$

Assume:

1. If $E(p) > 0$, buyers will increase their offer price.
2. If $E(p) < 0$, sellers will reduce their selling price.
3. If $E(p) = 0$, neither buyers nor sellers will adjust their prices and the market is in equilibrium.

These were the assumptions made by the nineteenth-century economist Leon Walras in his analysis of market stability.

Given these assumptions, market equilibrium will be stable if a price increase reduces excess demand. If this were not the case, a price increase, by increasing excess demand, would cause buyers to bid up prices even further, increasing excess demand still more, and no equilibrium would ever be reached.

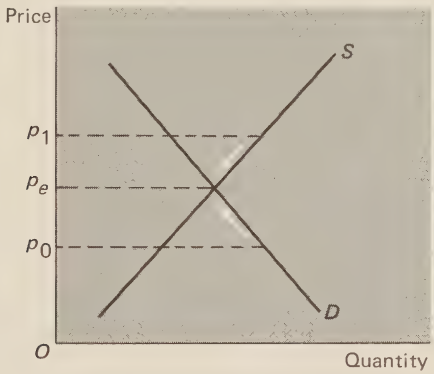


FIGURE 8–13
Stability Analysis of Market Equilibrium

Algebraically, the Walrasian stability condition can be stated as

$$\frac{\Delta E(p)}{\Delta p} = \frac{\Delta D(p)}{\Delta p} - \frac{\Delta S(p)}{\Delta p} < 0$$

or

$$\frac{\Delta D(p)}{\Delta p} < \frac{\Delta S(p)}{\Delta p}$$

Now, $\Delta D(p)/\Delta p$ is the change in quantity demanded divided by the change in price, which is the reciprocal of the slope of the demand curve.[12] Similarly, $\Delta S(p)/\Delta p$ is the reciprocal of the slope of the supply curve. Therefore, if the demand curve is negatively sloped and the supply curve is positively sloped, the market equilibrium will be stable.

This result can be seen graphically in Figure 8–13. Suppose the price is above

[12] Walras expressed his market demand and supply functions with quantity on the vertical axis and price on the horizontal axis. This specification makes more sense mathematically if we assume quantities demanded and supplied depend on price. Alfred Marshall introduced the use of price on the vertical axis and quantity on the horizontal axis, which is now the way market functions are conventionally represented in the economics literature.

equilibrium, at p_1. Since supply exceeds demand, excess demand is negative and price will be bid downward, toward p_e. If price is below p_e, at p_0, excess demand will be positive and price will be bid upward, toward p_e.

Instability in the classical analysis occurs when the demand and supply curves slope in the same direction. Say, for instance, that the supply and demand curves are both negatively sloped, as shown in Figure 8–14. This could be, for instance, the negatively sloped portion of the market supply curve for labor. If the demand curve is steeper than the supply curve, then

$$\frac{\Delta D(p)}{\Delta p} > \frac{\Delta S(p)}{\Delta p}$$

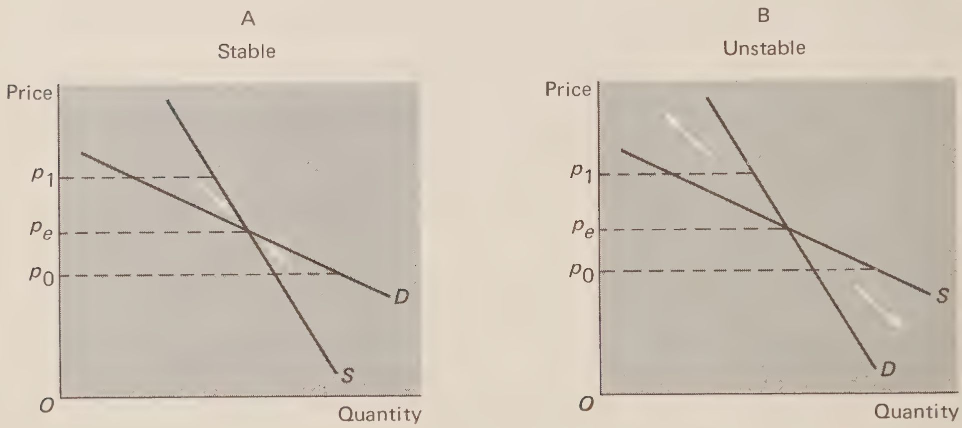


FIGURE 8–14
Walrasian Stability Analysis

and the equilibrium position at p_e will be unstable. In part B of Figure 8–14, when price is above p_e, at p_1, excess demand is positive, so prices will be bid upward, away from equilibrium. When price is below p_e, at p_0, excess demand is negative and price will be bid downward, again away from p_e. On the other hand, if the supply curve is steeper than the demand curve, then

$$\frac{\Delta D(p)}{\Delta p} < \frac{\Delta S(p)}{\Delta p}$$

and the equilibrium is stable, as shown in part A of Figure 8–14.

The reader should verify that if both the demand and supply curves are positively sloped, stability is achieved only if the demand curve is steeper than the supply curve. Also, in the unlikely event of a positively sloped demand curve and a negatively sloped supply curve, the equilibrium is unstable.

The Marshallian Approach

Suppose we make a different assumption about market behavior. The Walrasian model focused on price adjustments in response to excess demand or supply. Alfred Marshall, who was also interested in market stability, concentrated on adjustments in quantity demanded and quantity supplied in response to discrepancies in buying and selling prices. Marshall defined excess demand price, $F(q)$, as the difference between the price that buyers are willing to pay for some quantity q of the good, $D^{-1}(q)$,[13] and the price at which sellers are offering that amount, $S^{-1}(q)$, that is,

$$F(q) = D^{-1}(q) - S^{-1}(q)$$

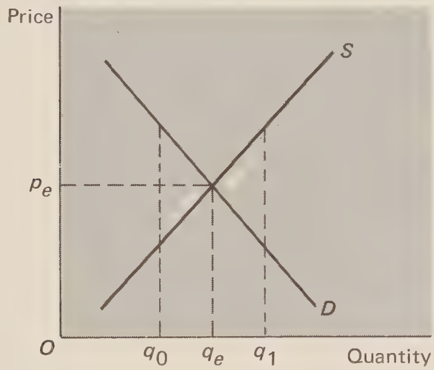


FIGURE 8–15
Marshallian Stability Analysis of a Normal Market

Marshall assumed:

1. If $F(q) > 0$, producers will increase quantity supplied.
2. If $F(q) < 0$, producers will reduce quantity supplied.
3. If $F(q) = 0$, sellers will not adjust quantity supplied and the market is in equilibrium.

Given these assumptions, market equilibrium will be stable if an increase in supply reduces excess demand price. If this were not the case, an increase in supply would increase excess demand price, causing sellers to increase supply even more. Equilibrium would never be reached.

Algebraically, the Marshallian stability condition can be stated as

$$\frac{\Delta F(q)}{\Delta q} = \frac{\Delta D^{-1}(q)}{\Delta q} - \frac{\Delta S^{-1}(q)}{\Delta q}$$

[13] Price as a function of demand, $D^{-1}(q)$, is the *inverse* of the quantity demanded as a function of price, $D(q)$.

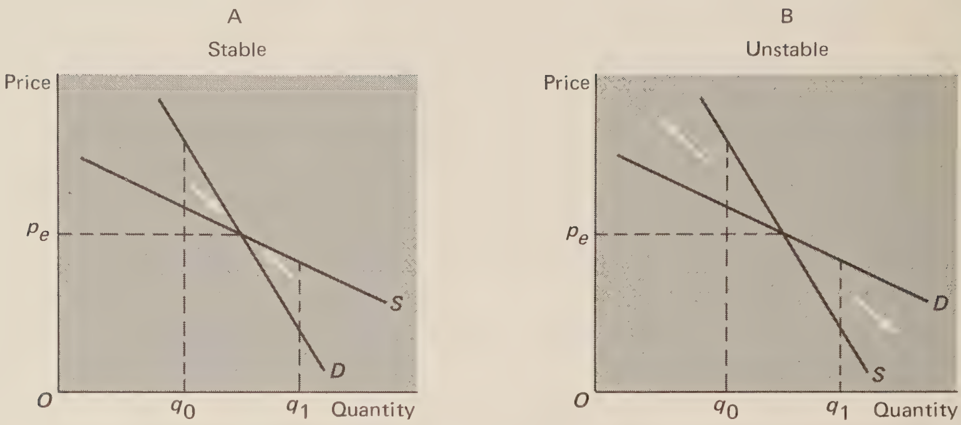


FIGURE 8–16
Marshallian Stability Analysis with Negatively Sloped Supply Curves

or

$$\frac{\Delta D^{-1}(q)}{\Delta q} < \frac{\Delta S^{-1}(q)}{\Delta q}$$

$\Delta D^{-1}(q)/\Delta q$ is the change in price divided by the change in quantity demanded, which is the slope of the demand curve. Similarly, $\Delta S^{-1}(q)/\Delta q$ is the slope of the supply curve. Therefore, if the supply curve is positively sloped and the demand curve is negatively sloped, the market equilibrium is stable.

Consider the graph in Figure 8–15. If output is below equilibrium, at q_0,

TABLE 8–1
Walrasian and Marshallian Stability Conditions *

	Walras	Marshall
S positive *D* negative	stable	stable
S and *D* negative *S* steeper	stable	unstable
S and *D* negative *D* steeper	unstable	stable
S and *D* positive *S* steeper	unstable	stable
S and *D* positive *D* steeper	stable	unstable
S negative *D* positive	unstable	unstable

* Supply and demand curves have price on the vertical axis and quantity on the horizontal.

excess demand price is positive and firms will increase supply toward q_e. If output is above q_e, at q_1, excess demand price is negative and producers will reduce quantity supplied.

As in the Walrasian model, instability will occur when the supply and demand curves slope in the same direction. However, the Marshallian stability conditions are exactly opposite from the Walrasian. As seen in Figure 8–16, the market equilibrium is stable when both curves are negatively sloped and the demand curve is steeper than the supply curve. It is unstable when the supply curve is steeper than the demand curve.

Table 8–1 summarizes the conditions for stability of market equilibrium under different assumptions about the slopes of the demand and supply curves as well as about the way the market participants adjust to disequilibrium situations. An analysis of stability in any particular market depends upon both types of considerations.

The Cobweb Model: Lagged Adjustment

The classical stability analyses of Walras and Marshall assumed that adjustment to disequilibrium situations was instantaneous. Although it was recognized that adjustment takes time, the time element in itself was not a crucial factor in explaining stability or instability of equilibrium.

If price adjustments to disequilibrium do not occur simultaneously with adjustments in quantities demanded or supplied, there is an additional source of instability. Models of market behavior under these conditions are called **cobweb models** because the time path of adjustment to equilibrium follows a cobweb pattern.

The example of price instability in the market for meat exhibited a cobweb effect. Suppose there is a shift in the supply function for meat because of the discovery of a new vaccine that reduces the incidence of a particular disease. Also assume that consumers are slow to adjust their purchases to price changes. Then, when the supply curve shifts from S to S', price falls from p_e to p_1, as shown in Figure 8–17. The decline in price eventually causes an increase in quantity demanded, from q_e to q_2. At q_1 suppliers will bid price up to p_2. The process of adjustment continues as indicated by the arrows, with price fluctuations increasing in amplitude and moving further and further away from the equilibrium price, p_e'.

The type of behavior described in this example does not necessarily lead to instability. In Figure 8–18 a shift in the supply curve from S to S' sets off an adjustment process in which price fluctuations become smaller and smaller until they finally converge to equilibrium. Initially, price declines from p_e to p_1 in response to the increase in supply. However, the resultant increase in quantity demanded, from q_e to q_1, is much smaller than in Figure 8–17. Price rises to p_2, but p_2 is below the previous equilibrium price. Output and price adjustments

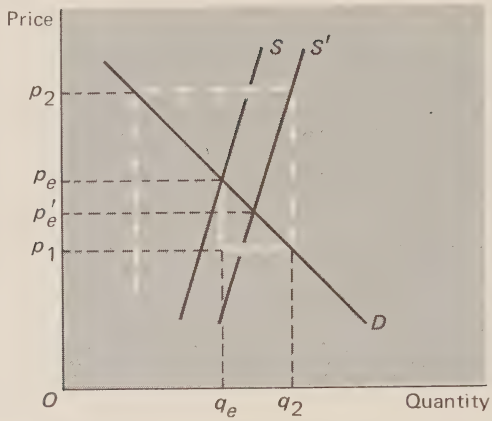


FIGURE 8–17
Cobweb Effect in the Meat Market

become smaller and smaller, until eventually the new equilibrium, p_e', is reached, as indicated by the arrows.

The difference between the markets in Figure 8–17 and Figure 8–18 is the relative slopes of the demand and supply curves. When the demand curve is flatter than the supply curve the new equilibrium will be unstable. When the demand curve is steeper than the supply curve the adjustment process will converge to equilibrium.

There are many different types of dynamic behavior that produce cobweb effects. They all involve a lag in adjustment of one of the variables, price or

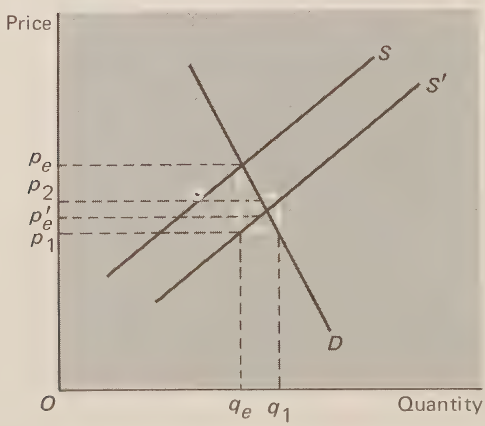


FIGURE 8–18
Convergence to Equilibrium in the Meat Market

quantity, to the other. Suppose, for instance, there is an increase in the demand for meat. If there is a lag in the response of quantity supplied to changes in market price, a cobweb effect will occur. The reader should verify that in this case stability requires that the demand curve be *flatter* than the supply curve.

Because of the length of the growing period, we normally expect supply lags in most agricultural markets. This explains why there are such great fluctuations in agricultural prices. This is a serious problem for farmers, particularly in underdeveloped countries, where the bulk of the population is engaged in farming. Since demand for most farm products is inelastic, price fluctuations generate wide swings in income. International agreements to stabilize agricultural prices by buying and selling stockpiles of commodities have helped to alleviate the problems to some extent. However, since agricultural goods are perishable, such programs are very costly when prices are depressed for long periods of time.

Certain factor markets are also subject to cobweb effects. When long training periods are required, the supply of labor in these occupations will lag in response to wage changes. The market for scientists and engineers that we discussed in chapter 7 has demonstrated a cobweb effect. In this market an increase in demand had the impact of raising wages, which attracted more people into the field. The long training period effectively restricted supply for a number of years, maintaining high wages and attracting even more people. By the time wages began to fall, there were still a large number of people undergoing training. The decline in demand in the late 1960s only served to intensify oversupply.

8–5 GENERAL EQUILIBRIUM

In the preceding discussion we examined market equilibrium simply in terms of the demand and supply functions in that market. The stability analysis also focused on the way participants in that market adjust to disequilibrium situations.

However, in an actual economy, markets for different goods and services are interrelated. We cannot examine equilibrium in one market if changes in demand and supply affect demand and supply conditions in other markets, which then feed back to the original market. One example of the interrelation between markets is the corn-hog relation. Corn is the major feed for hogs. Suppose the demand for hogs rises. Not only will hog prices rise but there will be an increase in the demand for corn, which will raise corn prices. The increase in corn prices will increase production costs for hog producers, shifting the supply curve leftward and causing a further increase in hog prices. Thus an increase in demand for hogs will eventually cause a shift in the supply function for hogs because of the effect on corn prices. If the hog market were studied in

isolation, the increase in hog prices associated with the increase in demand would be underestimated.[14]

An analysis of equilibrium in a single market is called a **partial equilibrium analysis.** Recognition of the interdependence of markets and an examination of all feedback effects is called a **general equilibrium analysis.** Whether we use a partial or a general equilibrium approach may depend on the degree of isolation of the particular market from the rest of the economy. For instance, if a single small firm undertakes a technological change that increases labor productivity, the demand for labor in that firm is likely to increase, since the impact on total market output, and hence on product price, will be negligible. On the other hand, if the change is adopted by the entire industry, the increase in labor productivity may be more than offset by a decline in product price, particularly if the demand for the product is inelastic. Consequently, the impact on market demand is not simply the horizontal sum of the firms' demand curves. The impact of the change in industry output must be taken into account in a general equilibrium approach.

General equilibrium analysis is important for other reasons too. Since there are many interrelated goods and factor markets, it is possible that equilibrium cannot be attained in all of them simultaneously. This is an important question for any economic system concerned with efficient resource allocation. Does there exist a set of positive prices that will clear all goods and factor markets simultaneously? An associated question is that of stability. How must market participants behave in disequilibrium situations if the general equilibrium is to be stable, that is, if it is ever to be achieved?

Finally, a related question concerns the consistency of the general equilibrium with the conditions for Pareto efficiency. If all the marginal conditions are met, will the general equilibrium be attained? Will all markets be cleared when production occurs where the marginal rate of substitution between any two goods or factors of production is equal to their marginal rate of transformation?

Input-Output Analysis

Although the earliest models of general equilibrium were developed by economists like Walras and Pareto in the nineteenth century, a more modern version has been applied extensively to practical problems of economic planning. The **input-output analysis** introduced by Wassily Leontief in the 1930s is used for planning by the USSR and other socialist countries as well as by large-scale businesses and governments in capitalist economies.

[14] If hog prices rise when the supply function shifts leftward, the quantity demanded will decline, reducing the demand for corn and causing corn prices to fall. This will cause a rightward shift in the supply curve of hogs, reducing the price, which will increase quantity demanded. Whether the process ever results in a stable equilibrium price in both markets depends upon the relative slopes of the supply and demand curves in each market. For a discussion see James M. Henderson and Richard E. Quandt, *Microeconomic Theory,* 2d ed. (New York: McGraw-Hill, 1971), pp. 145–149.

Input-output analysis focuses on the production side of the economy. Leontief recognized that many products are inputs in other production processes. Hence, to obtain any amount of goods and services for final consumption a larger amount of goods and services must initially be produced. The difference between output and final consumption of a single product will depend upon the need for that product for producing other goods. If we have an estimate of final demand for goods and services for consumption and investment, and by government, input-output analysis can be used to determine production requirements for each industry. On the other hand, if we know the production capabilities of the economy, we can project the amount of goods and services that will be available for final sales.

Consider an economy producing both manufactured and agricultural goods. Agricultural raw materials are used in manufacturing processes. For every dollar of manufacturing output, 20 cents worth of agricultural raw materials is required. In addition, 10 cents worth of agricultural raw materials is used in agricultural production. Farm equipment comes from manufacturing industry. Thirty cents worth of manufacturing equipment is used for every dollar's worth of agricultural output. Manufacturing industry uses 40 cents worth of manufactured equipment to produce a dollar of output. These relations between the various industries are shown on the input-output table, Table 8–2. Notice that they are in terms of dollars worth of input and output rather than physical units. Since each dollar of agricultural output requires 10 cents worth of agricultural products and 30 cents of manufactured products, the remainder of the value, 60 cents, is inputed to labor, the only nonproduced factor of production.

TABLE 8–2
Input-Output Table

	OUTPUT (Dollars)	
INPUT (Dollars)	Agriculture	Manufacturing
Agriculture	0.1	0.2
Manufacturing	0.3	0.4
Labor	0.6	0.4
TOTAL	1.0	1.0

Suppose we wish to produce \$100 million of agricultural output and \$200 million of manufactured goods for final consumption. Output of agricultural goods must include allowance for direct production requirements. Therefore, total agricultural output, A, must be

$$A = 0.1A + 0.2M + \$100 \text{ million}$$

where M is output of manufactured goods. Similarly,

$$M = 0.4M + 0.3A + \$200 \text{ million}$$

Here we have two unknowns, A and M, and two equations that can be solved simultaneously to give

$$A = \$208.3 \text{ million}$$
$$M = \$437.5 \text{ million}$$

To obtain \$100 million of agricultural output and \$200 million of manufactured output for final consumption \$208.3 million of agricultural output and \$437.5 million of manufactured output must be produced. Notice that both A and M must be produced simultaneously, if either one is produced at all.

Suppose there are 3 million man-hours of labor in the labor force. If \$208.3 million worth of agricultural output is produced, then 0.6(\$208.3 million), or \$125 million, will go to labor in wages. In manufacturing, 0.4(\$437.5 million), or \$175 million, will be available for wages. An hourly wage of \$1 will ensure full employment of labor, so 125 million workers will be employed in agriculture and 175 million in manufacturing. Furthermore, the total amount of wages, \$300 million, will be exactly sufficient to purchase the products available for final consumption.

Limitations of the Input-Output Approach

The input-output approach focuses on the supply side of market. Consumption or final demand is taken as given, and prices are assumed to be set to clear product markets. This is an important limitation of the model, since there is no mechanism to ensure that product markets will be cleared. In socialist countries, as has been mentioned, forecasting consumer demand and establishing prices that will clear product markets has been the most serious problem for economic planning. Neither the Lange model of efficient resource allocation under socialism nor actual experience has provided a solution. Input-output does not address itself to this issue.

As a model of the interdependence of industries in production, the Leontief approach also has limitations. It is assumed that interindustry requirements are fixed, so that there is only one level of output for each industry consistent with the stipulated requirements for final demand. If one industry does not meet its targets, output in all other industries must decline. If there were alternative techniques for producing goods, shortages in one industry could be made up by surpluses in another.

Despite these limitations, input-output analysis is a powerful tool for analyzing interrelations within an economy. Suppose an underdeveloped country is considering encouragement of a domestic steel industry. Input-output analysis can be used to project the demands for goods and services to be used in the production of steel. This kind of analysis is also important for determining import requirements when foreign exchange is scarce. An input-output table can be used to pinpoint those industries that are the greatest users of imports. Input-output can

take into account indirect as well as direct resource requirements. Agriculture, for instance, may use imports indirectly through its use of manufactured equipment.

The Walrasian General Equilibrium System

The models of general equilibrium developed by the nineteenth-century economists were more comprehensive than the input-output model. What they offered in generality, however, they lost in usefulness. Walras developed a set of equations to describe the operation of a market system. However, the functional form of the equations is left unspecified, and consequently the system is not very useful for actual planning purposes. It does, however, provide proof of the existence of general equilibrium, its consistency with Pareto efficiency, and a demonstration of the conditions for stability.

Let us begin with an economy in which m goods and services are traded among n individuals. Each individual (referenced by subscript i) holds $\overline{X}_{ri}$, an amount of commodity r, at the outset and trades with other individuals, holding X_{ri} of the rth commodity after exchange. There is no new production of any of the commodities during the period. The price of the rth good is p_r.

Each household will exchange goods until its utility is maximized. This implies that for the ith household,

$$\frac{MU_i}{p_i} = \cdots = \frac{MU_r}{p_r} = \cdots = \frac{MU_m}{p_m} \tag{1}$$

In addition, the budget constraint must be satisfied so that

$$\sum_{r=1}^{m} p_r(X_{ri} - \overline{X}_{ri}) = 0 \tag{2}$$

The notation $\sum_{r=1}^{m}$ means that this is a sum of the expression for all m commodities.

Finally, to achieve general equilibrium the market for all m goods must be cleared. This means that the quantity of each good held by all individuals taken together after trade must equal the amount held before trade. This implies

$$\sum_{i=1}^{n} (X_{ri} - \overline{X}_{ri}) = 0 \tag{3}$$

In this case, the expression is a sum of the quantities of commodity r held before and after trade by all n households.

A general equilibrium exists in this market if these conditions provide enough equations to solve for the m equilibrium prices, $p_1, \ldots, p_m$, of the m goods and mn quantities, $X_{11}, \ldots, X_{1n}, \ldots, X_{m1}, \ldots, X_{mn}$, of the m goods held by the n households.[15]

[15] Additional criteria are necessary to demonstrate that these prices and quantities are positive. Furthermore, the equality of equations and variables does not always ensure a unique solution or any solution at all. For an advanced mathematical discussion and proof of existence see Kenneth J. Arrow and Gerhard Debreu, "Existence of an Equilibrium for a Competitive Economy," *Econometrica*, *XXII* (July 1954), pp. 265–290.

In the first group (1), there are $m-1$ equations for each of the n households, a total of $n(m-1)$. In (2) there are n budget equations, one for each household. Finally, in (3) there are m market-clearing equations, one for each good.

Walras showed that if all the consumers' budget constraints are satisfied and if $m-1$ markets are cleared, then the mth market will also be cleared.[16] This result, known as **Walras's Law,** means that the general equilibrium can be achieved with only $m-1$ equations in (3). The total number of equations in (1), (2), and (3) is therefore $n(m-1)+n+(m-1)$, or $mn+(m-1)$. Therefore, we cannot solve for mn quantities and m prices, but only for mn quantities and $m-1$ prices. The price of one good is arbitrary, that is, it can take on any value. General equilibrium exists only for a set of relative prices and the absolute price level is arbitrary.[17]

This is, of course, consistent with the theory of resource allocation. All values are relative concepts. The marginal rate of substitution is the preference for one good in terms of another. The marginal rate of transformation is the opportunity

---

[16] The proof is as follows: If the condition

$$\sum_{i=1}^{n} (X_{ri} - \overline{X}_{ri}) = 0$$

holds for $m-1$ markets, and if the condition

$$\sum_{r=1}^{m} p_r(X_{ri} - \overline{X}_{ri}) = 0$$

holds for all n consumers, so that

$$\sum_{i=1}^{n}\sum_{r=1}^{m} p_r(X_{ri} - \overline{X}_{ri}) = \sum_{r=1}^{m} p_r\left[\sum_{i=1}^{n} (X_{ri} - \overline{X}_{ri})\right] = 0$$

then subtracting, we get

$$\sum_{r=1}^{m} p_r\left[\sum_{i=1}^{n} (X_{ri} - \overline{X}_{ri})\right] - \sum_{r=1}^{m-1} p_r\left[\sum_{i=1}^{n} (X_{ri} - \overline{X}_{ri})\right] = 0$$

or

$$p_m \sum_{i=1}^{n} (X_{mi} - \overline{X}_{mi}) = 0$$

Assuming p_m is not zero, this implies

$$\sum_{i=1}^{n} (X_{mi} - \overline{X}_{mi}) = 0$$

so the mth market is cleared, which was what was to be proved.

[17] If money has a value other than as a medium of exchange, then money can be treated as an additional commodity and an absolute level of prices (in terms of the monetary unit) can be determined. Keynes suggested in his theory of the speculative demand for money that money may be useful as a store of value. For a discussion of the impact of this assumption for prices in a Walrasian general equilibrium system see Oskar Lange, "Say's Law: A Restatement and Criticism," in Lange, McIntyre and Yntema (eds.), *Studies in Mathematical Economics and Econometrics* (Chicago: University of Chicago Press, 1942), pp. 49–68; and Don Patinkin, *Money, Interest, and Prices,* 2d ed. (New York: Harper and Row, 1965), chapter 3.

cost of one good in terms of another. The equilibrium value equates the marginal rate of substitution with the marginal rate of transformation, so that we can only conceive of the value of a good or resource in terms of another good or resource.

Extensions of the Walrasian Model

Suppose we allow for production to take place. Assume that k of the m commodities are factors of production (represented by subscript s) and $m - k$ commodities are consumer goods (represented by subscript t). For simplicity, also assume that the technical coefficients of production are fixed, so that a_{st} is the amount of factor s used to produce a unit of output of t. Let Y_r refer to the output of the commodity in new production.

Now, we cannot use more of a factor than is made available over the period. Thus total factor use must equal the total amount available, or

(4) $$\sum_{t=m-k}^{m} a_{st} Y_t = -Y_s$$

Finally, we assume that the firm sets price equal to marginal cost, either as a profit maximizer or a regulated Lange enterprise. Since there are constant returns to scale, marginal cost is equal to average cost, which is the cost of the inputs for each unit of output, that is,

$$MC = AC = \sum_{s=1}^{k} p_s a_{st}$$

Therefore, when price is equal to marginal cost,

(5) $$\sum_{s=1}^{k} p_s a_{st} - p_t = 0$$

In addition, the market-clearing equations in (3) must be expanded by the amount of new production, that is,

(3′) $$\sum_{i=1}^{n} (X_{ri} - \overline{X}_{ri}) = Y_r$$

Now we have added m new variables to our system, $Y_1, \ldots, Y_m$, the new production of commodities. In (3′) we have changed the equations in (3) but have not increased their number. There are k new equations in (4), one for each factor of production. In (5) there are $m - k$ new equations, one for each consumer good. Thus we have added m new equations to our system, which will allow us to determine equilibrium values for the m new variables.

Limitations of the Walrasian Approach

Economists since Walras have pointed out that having the same number of equations as variables does not always ensure a unique solution or, for that matter, any solution at all. For instance, the system of equations

$$x^2 + y^2 = 0$$
$$x^2 - y^2 = 1$$

has no real solution. Furthermore, the solution may not be economically meaningful in the sense that the equilibrium might occur at negative prices and outputs. Kenneth Arrow and Gerhard Debreu have investigated the conditions under which the Walrasian system does possess an economically meaningful (but not necessarily unique) real solution. These conditions, known as the **convexity conditions** include nonincreasing returns to scale and the absence of externalities in production or consumption.[18]

An intuitive explanation of the convexity conditions is that they imply scarcity of resources and satiability of wants. The opportunity set must be convex from above, that is, opportunity costs must increase with the output levels of goods. The utility set must be convex from below, implying that the marginal rate of substitution declines the more of a good is consumed.

These conditions have analogs in partial equilibrium analysis. Diminishing returns and diminishing marginal rates of transformation reflect scarcity of resources. They are sufficient for the marginal conditions to produce Pareto efficiency in production. Diminishing marginal utility and diminishing marginal rates of substitution between commodities reflect satiability of wants. These are sufficient for the marginal conditions to produce Pareto efficiency in consumption. Thus the convexity conditions required to prove the existence of general equilibrium are also the sufficient conditions for the attainment of Pareto efficiency.

The Walrasian general equilibrium is also consistent with the attainment of the Paretian marginal conditions. We have assumed that consumers set the marginal rate of substitution between commodities equal to the ratio of their prices. Producers expand production until marginal cost equals price, so that the marginal rate of transformation between commodities is equal to the ratio of their prices. Since all commodities are traded in markets, and the markets are cleared, the marginal rate of substitution between commodities in consumption is equal to their marginal rate of transformation in production. Thus the Paretian conditions are established.[19]

---

[18] An excellent but difficult discussion of general equilibrium, its existence, and its relation to the Paretian conditions is James Quirk and Rubin Saposnik, *Introduction to General Equilibrium Theory and Welfare Economics* (New York: McGraw-Hill, 1968), chapters 3 and 4.

[19] We could drop the assumption of fixed technical coefficients of production. If we include k equations stating the proportionality of the marginal products of the factors to their prices for each of the $m - k$ commodities, we would get $k(m - k)$ new equations that would determine $k(m - k)$ new input coefficients. These would satisfy the Paretian conditions for optimal factor utilization.

The existence of general equilibrium consistent with optimal resource allocation in the presence of externalities and increasing returns to scale is an important issue in the theory of resource allocation. In the case of externalities, it is generally assumed that government or some outside agent must intervene in the general market equilibrium if efficiency is to be achieved. The impact of such intervention on multimarket equilibrium is a subject that deserves considerably more attention than it has received in the literature.

If returns to scale are increasing, marginal cost pricing involves losses that, in a free enterprise economy, must be reimbursed if the firm is to remain in business. In the presence of economies of scale firms can undercut their competitors by expanding in size and reducing unit costs of production. In this case, firms become large enough to monopolize or dominate the market. Price is no longer viewed as fixed by external market forces, and the model of marginal cost pricing is no longer applicable. We will consider the question of equilibrium under alternative market structures in chapter 9.

Stability of General Equilibrium

Stability analysis is as important for general equilibrium as it was for the use of the equilibrium concept in a single market. In general equilibrium, stability analysis must take into account the impact of price adjustments in interrelated markets and their feedbacks on one another. Although the analysis of multimarket stability involves sophisticated mathematical techniques that are beyond the scope of this text, the approach is similar to the analysis of stability in a single market.[20]

QUESTIONS FOR STUDY AND REVIEW

1. The input-output matrix for a three-sector economy is as follows:

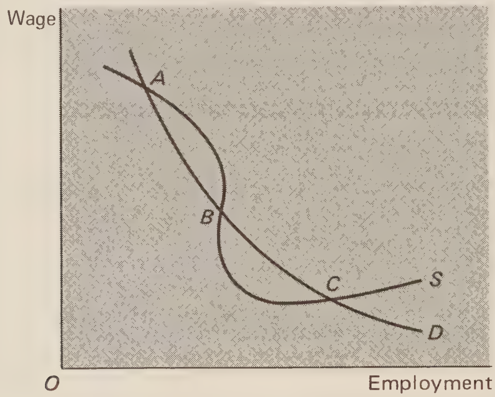
	OUTPUT		
INPUT	Agriculture	Manufacturing	Services
Agriculture	0	0.4	0.3
Manufacturing	0.2	0	0.5
Services	0.1	0.2	0
Labor	0.7	0.4	0.2

a) If each sector has a maximum capacity of $1 billion per year, what is the maximum output each sector can produce for final consumption?

b) If the labor force consists of 1 billion man-hours per year, what will be the equilibrium wage? How will employment be distributed among the sectors of the economy?

[20] For a discussion of stability in general equilibrium see Quirk and Saposnik, *General Equilibrium Theory*, chapter 5; and Lloyd A. Metzler, "Stability of Multiple Markets: The Hicks Conditions," *Econometrica, XIII* (October 1945), pp. 277–292.

2. The demand and supply functions for labor in a particular country are as follows:



Discuss the stability conditions for the equilibrium points, *A, B,* and *C*.

3. "Although equilibrium itself is a static concept, it is only useful in the context of a dynamic theory." Discuss, illustrating your argument with hypothetical examples.
4. Consider a market in which there is a one-period lag in supply and no demand lag. What restrictions on the supply and demand functions are required for this market to be stable?
5. Consider an economy with three goods, two consumers, and two factors of production. Set up a general equilibrium model for this economy, demonstrating its consistency with the attainment of Pareto efficiency.

ADDITIONAL READING

Allen, R. G. D. *Mathematical Economics,* 2d ed. London: Macmillan and Co., Ltd., 1963, chapters 1, 10, and 11.

Baumol, William J. *Economic Dynamics,* 3d ed. New York: Macmillan, 1970, chapter 7.

Buchanan, Norman S. "A Reconsideration of the Cobweb Theorem," *Journal of Political Economy, XLVII* (February 1939), pp. 67–81.

Dorfman, Robert. "The Nature and Significance of Input-Output," *Review of Economics and Statistics, XXXVI* (May 1954), pp. 121–133.

Kuenne, Robert E. *The Theory of General Equilibrium.* Princeton: Princeton University Press, 1963, chapter 1.

Leontief, Wassily. *The Structure of the American Economy, 1919–1939,* 2d ed. New York: Oxford University Press, 1951, part III.

Quirk, James, and Rubin Saposnik, *Introduction to General Equilibrium Theory and Welfare Economics.* New York: McGraw-Hill, 1968.

Samuelson, Paul A. *Foundations of Economic Analysis.* Cambridge, Mass.: Harvard University Press, 1947, chapters 1, 2, and 9.

Walras, Leon. *Elements of Pure Economics.* Translated by William Jaffe. London: Allen and Unwin, 1954, lessons 11 and 12.

part III

BEYOND THE STATIC COMPETITIVE MODEL

9 Market Structure and Performance

In the preceding chapters we developed a model of production and market supply that we have called the market model. We have assumed that the output of each firm or producer is so small in relation to total industry output that its actions have a negligible effect on market price. The firm can sell all it wants to at the market price and the impact on total market supply will not be noticed. Under these conditions firms maximize profits where price is equal to marginal cost, and therefore profit maximization by individual firms produces Pareto efficiency for the producing sector as a whole, assuming no external effects. For the producing sector, this is what Adam Smith meant when he argued that if each individual seeks his own best interests, the interest of society will be served.

This particular model of the firm is called the market model because this type of market structure will produce Pareto efficiency in a market economy. But even in the absence of externalities, when firms are so large that their operations have a significant effect on market supply, profit maximization no longer produces Pareto efficiency in most cases. Although a few industries in the United States economy consist of many small producers as required by the market model—agriculture, some types of retail trade—the majority of American manufacturing industries are dominated by a few large firms. It is important to recognize that the market model is a description of an ideal type of market structure, under static conditions, rather than of the way the United States economy (or any other economy, for that matter) actually works.[1]

[1] A distinction between "ideal" as a normative model and "ideal type" must be made. To the extent that Pareto efficiency has normative consequences, the market model is an "ideal." However, as was noted, distributive issues limit the normative implications of Pareto efficiency. "Ideal type" merely refers to a conceptual system that portrays certain hypothetical events. Ideal types are not intended to be descriptions of reality. Max Weber, a sociologist who developed the concept of the ideal type states that "its relationship to the empirical data consists solely in the fact that where market-conditioned relationships of the type referred to by the abstract construct are discovered or suspected to exist in reality to some extent, we can make the *characteristic* features of this relationship pragmatically *clear* and *understandable* by reference to the *ideal-type*." Max Weber, "Ideal Types and Theory Construction," in May Brodbeck (ed.), *Readings in the Philosophy of the Social Sciences* (New York: Macmillan, 1968), p. 497.

9–1 THE STRUCTURE OF AMERICAN INDUSTRY

Table 9–1 shows the per cent of total sales, assets, and profits made by the four largest firms in the major manufacturing industries in the United States in 1962.

TABLE 9–1
Concentration of Sales, Total Assets, Net Capital Assets, and Profits After Taxes for the Four Largest Firms in Twenty-Eight Selected Industry Groups, Fourth Quarter, 1962.

	Per Cent of Total			
Industry	Sales	Total Assets	Net Capital Assets	Profits
Motor Vehicles	80.8	79.7	83.1	89.1
Tobacco	70.9	72.7	69.8	72.5
Petroleum Refining	50.3	50.1	47.7	54.3
Rubber	48.1	55.0	56.4	51.6
Aircraft	47.3	41.9	32.6	46.6
Dairy Products	42.9	48.8	47.4	73.9
Basic Industrial Chemicals	42.0	45.5	46.6	64.6
Alcoholic Beverages	41.4	47.2	30.8	58.3
Primary Iron and Steel	40.2	48.0	48.8	44.3
Instruments	37.9	41.2	50.2	56.6
Electrical Machinery	34.3	35.6	41.5	44.4
Bakery Products	33.6	39.6	38.2	52.8
Drugs and Medicines	31.0	29.2	33.3	32.6
Other Transportation Equipment (Except Motor Vehicles and Aircraft)	30.3	44.2	59.9	51.6
Other Chemicals (Except Basic Industrial and Drugs and Medicines)	28.5	30.0	33.6	35.8
Primary Nonferrous Metals	27.3	41.1	47.7	37.1
Leather	26.7	32.1	35.4	28.8
Textile Mill Products	22.0	26.1	25.7	30.5
Lumber and Wood Products	21.2	31.0	41.5	48.1
Paper	20.7	23.2	22.3	35.0
Other Machinery (Except Electrical and Metalworking)	20.6	24.3	31.5	39.6
Stone, Clay, and Glass Products	18.1	19.1	19.8	23.4
Miscellaneous Manufacturing	16.3	33.1	34.3	25.2
Other Fabricated Metal Products (Except Primary Metals)	14.7	19.9	30.3	17.7
Metalworking Machinery	14.5	16.3	18.5	19.1
Other Food (Except Dairy and Bakery Products)	12.5	13.2	14.9	20.1
Furniture and Fixtures	5.2	8.4	9.6	5.3
Apparel	4.9	7.7	11.4	7.4

Source: Bureau of Economics, Federal Trade Commission.

Of the twenty-eight listed, fourteen industries had four firms that made over 30 per cent of industry sales, and in twenty industries, four firms made over 30 per cent of the profits. Notice that in only one case, aircraft, was the profit share of the largest four firms less than their share of sales. This implies that the dominant firms in the industry have on the average a higher profit to sales ratio than the other firms in the industry.

In 1962 there were 420,000 firms in American manufacturing industry, yet the twenty largest accounted for 25 per cent of their total assets, the fifty largest for 35.7 per cent, the hundred largest for 46.1 per cent, and the thousand largest for 74.8 per cent. This means that the 419,000 smallest companies accounted for only 25.2 per cent, approximately the same as the twenty largest.[2]

When sales of each industry are weighted by its **value added,** the difference between its sales and purchases, the average per cent of sales accounted for by the four largest firms is 37 per cent.[3] For the average American industry, therefore, the actions of the dominant firms will have a significant impact on market supply.

9–2 DETERMINANTS OF INDUSTRIAL CONCENTRATION

Since the market model of many small firms produces a theoretically efficient allocation of resources, it is important to consider how a high degree of concentration affects market performance. An industry is said to be highly concentrated if a few firms account for a large share of sales. Table 9–1 shows that there is considerable variation among American industries. Motor vehicles and tobacco are highly concentrated, while in the furniture and apparel industries the four largest firms account for only about 5 per cent of sales.

Antitrust legislation, designed to improve the performance of American industry, has traditionally focused on reducing industrial concentration. In part, the policy is based on extrapolation from the market model. If many small firms produce efficient resource allocation, then it must always be a good policy to reduce the degree of industrial concentration. To evaluate such a proposition we need a theory that relates market performance to market structure. In addition, we must analyze the determinants of market structure to understand why there is so much variation in the first place.

Barriers to Entry

When we think of an industry dominated by a few firms, we typically conjure up visions of prosperous businessmen in a smoke-filled room plotting collusive

[2] From testimony by Willard E. Mueller, chief economist and director of the Bureau of Economics of the Federal Trade Commission, before a subcommittee of the Senate Judiciary Committee in 1965. Reprinted in Edwin Mansfield (ed.), *Monopoly Power and Economic Performance*, rev. ed., (New York: W. W. Norton, 1968), pp. 73–77.

[3] From testimony by Morris Adelman, professor of economics at Massachusetts Institute of Technology, before a subcommittee of the Senate Judiciary Committee in 1965. Reprinted in Mansfield, *Monopoly Power and Economic Performance*, p. 79.

price agreements and ways to undermine potential entrants into the industry.[4] However, although historical and institutional factors that have produced concentrated market control may explain high concentration in some industries, in many cases barriers to entry occur naturally and are due entirely to economic or technological considerations. In these cases, high degrees of industrial concentration can be maintained without overt collusion and without the use of unethical tactics on the part of the firms in the industry.

The degree of concentration in any industry depends upon barriers to the entry of new firms. Given the state of market demand, the industry will be more concentrated if fewer firms are able to enter. Barriers to entry fall into three categories: physical, institutional, and economic.

Physical Barriers

Physical barriers to entry are associated with the ownership of scarce resources that are necessary inputs to the product. Classical models of noncompetitive market structures generally alluded to physical barriers as the mechanism for preventing entry. Cournot, for example, based his analysis of duopoly (an industry composed of two firms) on the provision of water from two mineral springs, each exploited by a different owner. Petroleum and mining are examples of industries where physical barriers could prevent unlimited entry.

Institutional Barriers

We often think of institutional barriers as mechanisms that promote oligopolistic (a few firms) market structures. Patents and licensing can prevent new firms from entering an industry in its early stages of development. At later stages these factors limit the ability of some firms to offer a competitive product. Another important institutional barrier is the good will associated with products of established firms. Patents and licensing combined with good will can exert powerful barriers to entry. In the 1920s the first automobile manufacturers were protected by patents. Although the patents eventually expired, consumer good will, or "brand loyality" associated with the products of the established firms, made it extremely difficult for new firms to capture a significant share of the market.

Economic Barriers

Barriers to entry may be based on purely economic factors and would occur in the absence of physical limitations or special institutional arrangements. Since these barriers would exist in any economic system, we would expect the degree of

---

[4] This is not to suggest that such situations do not occur in the real world. See, for example, Richard A. Smith, *Corporations in Crisis* (New York: Doubleday, 1963), chapters 5 and 6.

concentration attributable to such barriers to be similar for similar industries in different countries. If, for example, high concentration in the aircraft industry is due to economic factors, we would expect the same degrees of concentration in different countries with markets of about the same size.

Economies of scale are an important barrier to the entry of new firms in an industry. Even if the existing firms are enjoying substantial profits, that is, price is above average cost, the optimal size of the firm may be so large that entry of an additional firm would drive market price below average cost. If indivisibilities and the need for specialized equipment make it profitable to produce only at

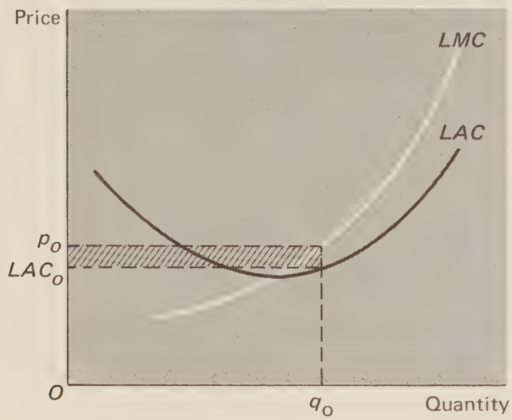


FIGURE 9–1
Entry-Inducing Price in the Market Model

very large scales, the impact on market supply of an additional entrant will be considerable. Potential entrants must consider the post-entry price rather than the pre-entry price to determine the profitability of undertaking a new enterprise.

Inelastic market demand also provides a barrier to entry. The more inelastic the demand the greater will be the impact of any given increase in supply on market price. Economies of scale and inelastic demand reinforce each other, permitting existing firms in the industry to earn rents, even in the long run. That is, they all can produce indefinitely where price exceeds long-run average cost.

To see why this is so, consider an industry in which all potential entrants have the same long-run cost functions as existing firms. There is no product differentiation, and no natural or institutional barriers to entry. In the market model firms would be attracted to the industry as long as firms in the industry are earning quasi-rents, that is, if price is above long-run average cost, *LAC*, as shown in Figure 9–1. The shaded area is the quasi-rent earned by existing firms. Since each firm is small in relation to the total market, entry will have a negligible effect on price, so that the individual firm will view the post-entry price as identical to the pre-entry price. In the market model, as many firms enter in re-

sponse to the quasi-rents, the market price will gradually fall in response to the increased supply. Entry will stop when price is equal to long-run average cost, as shown in Figure 9–2. Assuming *LAC* is U-shaped (since there are no significant economies of scale), the long-run equilibrium occurs at the plant size that minimizes *LAC* and where long-run marginal cost, *LMC,* is equal to *LAC*. Under the assumption that all firms have identical cost curves, there will be no rents or quasi-rents earned by the firms.[5]

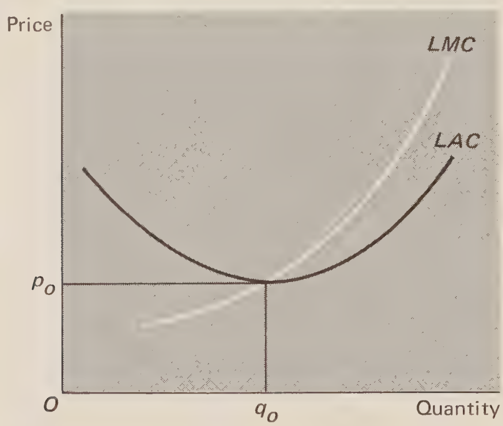


FIGURE 9–2
Entry-Preventing Price in the Market Model

The market model assumes that changes in supply due to entry are too small to affect market price, Suppose firms can enter only at very large scales relative to the size of the market. Even if the pre-entry price is above *LAC,* entry may drive price below minimum *LAC,* as shown in Figure 9–3. Conceivably, firms in the industry could keep prices at some level that would maintain an effective barrier to entry while they earn rents, or "monopoly profits." The amount of profits they can earn will depend on the magnitude of scale economies and the elasticity of demand for the product.[6]

Suppose existing firms wish to ascertain the highest entry-preventing price they can maintain in the long run, p_0. The associated market supply is q_0. Assume that $\overline{q}$ is the optimal scale of output and that output less than $\overline{q}$ for any firm would

[5] In most cases, some firms will have locational, managerial, or other advantages that will reduce their unit costs and allow them to earn rents. Only the least efficient firms in the industry will produce where price equals *LAC*.

[6] The analysis that follows was developed by P. Sylos Labini, *Oligopoly and Technical Progress,* rev. ed., trans. Elizabeth Henderson (Cambridge, Mass.: Harvard University Press, 1969), and Joe S. Bain, *Barriers to New Competition* (Cambridge, Mass.: Harvard University Press, 1956). An excellent summary is found in Franco Modigliani, "New Developments on the Oligopoly Front," *Journal of Political Economy, LXVI* (June 1958), pp. 215–232.

entail prohibitively high unit costs. If K is the minimum LAC associated with that plant size, then in the market model the price, p_c would be

$$p_c = K$$

with a corresponding output of q_c.

The size of the market, S, is the number of firms required to produce the competitive market output, that is,

$$S = \frac{q_c}{\overline{q}}$$

Entry-preventing output, q_o, is just enough to prevent another firm from entering without driving price below K, or

$$q_o = q_c - \overline{q} = q_c\left(1 - \frac{\overline{q}}{q_c}\right) = q_c\left(1 - \frac{1}{S}\right)$$

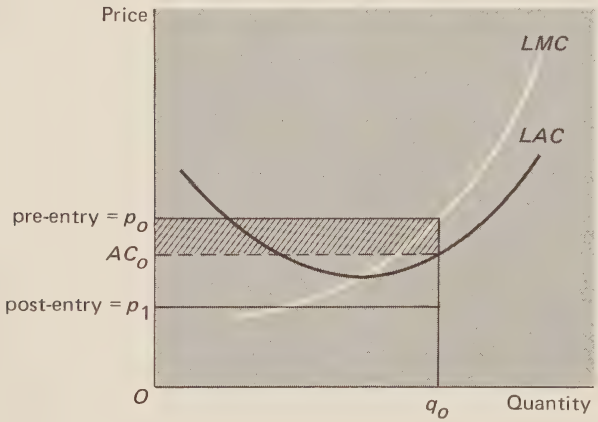


FIGURE 9-3
Entry-Preventing Price with Economies of Scale

The relation between the competitive market output (zero monopoly profits) and the entry-preventing output depends on the size of the market. Entry-preventing output will be $100/S$ per cent below output in the market model. If five firms would be required to produce the zero profit output, then entry-preventing output would be 20 per cent below q_c.

Since q_o is less than q_c, post-entry price, p_o, must be greater than minimum LAC. The actual amount of rents that existing firms can earn will depend on the elasticity of demand as well as on the size of the market. From the demand function, it can be shown that

$$p_o = p_c\left(1 + \frac{1}{ES}\right)$$

where E is the elasticity of demand.[7] Thus the post-entry price is $100/ES$ per cent above minimum LAC. If five firms would be required to produce the zero profit output, and demand elasticity is 2, the entry-preventing price would be 10 per cent above the competitive price, p_c. If demand elasticity were 1.5, p_o would be 17.5 per cent above p_c. The more inelastic the demand the greater the difference between the entry-preventing price and the competitive price, given the size of the market.

Characteristics of Concentrated Industries

When economic barriers to entry are considered, it is likely that capital-intensive industries will be among the most highly concentrated, since firms in such industries experience economies of scale (see chapter 6).[8] The more inelastic the market demand the greater will be the monopoly profits earned by existing firms as a result of economies of scale. Furthermore, inelastic demand is most likely to lead to collusion, that is, concerted efforts by existing firms to maintain price at an entry-preventing level. This is because firms operating in markets

---

[7] The proof is as follows:
Since

$$E = -\frac{p}{q} \times \frac{\Delta q}{\Delta p}$$

then

$$\frac{\Delta p}{p} = -\frac{\Delta q}{q} \times \frac{1}{E}$$

or

$$\frac{p_o - p_c}{p_c} = -\frac{q_o - q_c}{q_c} \times \frac{1}{E}$$

But

$$q_o - q_c = -\bar{q}$$

so that

$$\frac{p_o - p_c}{p_c} = \frac{\bar{q}}{q_c} \times \frac{1}{E} = \frac{1}{SE}$$

Rearranging terms,

$$p_o - p_c = \frac{p_c}{ES}$$

and

$$p_o = p_c\left(1 + \frac{1}{ES}\right)$$

[8] Economies of scale may be due to nontechnological considerations as well. The tobacco industry, for instance, has developed an extraordinary level of competitive advertising that requires such heavy fixed cost outlays that entry would only be profitable at large scales of operation. Thus highly concentrated industries are not necessarily capital-intensive, although the reverse is usually the case.

where demand is inelastic are most likely to recognize the impact of changes in market supply on market price. We shall see that overall price behavior in such markets is also affected by the recognition of the interdependence of firms' actions.

Consider for a moment some observable characteristics of highly concentrated industries. If they are capital-intensive, average labor productivity will be high. In many such industries special skills are required to operate the capital equipment and consequently wages will also be relatively high. Furthermore, in industries dominated by a few firms, labor has unionized to salvage a bargaining position vis-á-vis the large companies. In some cases, unions have successfully restricted the supply of labor possessing the requisite skills by imposing apprenticeships and other entrance requirements. Nevertheless, any evaluation of union claims that they can achieve higher wages for workers on the basis of past experience must take into account the fact that the highly concentrated, capital-intensive industries have the most effective unions. Although, in general, wages are higher in unionized than in nonunionized industries, this could be due to the nature of the industry rather than to the efforts of the unions per se.

Not only do wages tend to be higher in highly concentrated industries but average industry profits in the long run are likely to be higher too, given the ability of the firms to earn quasi-rents. This does not mean that profits in a concentrated industry cannot be seriously depressed because of short-run changes in demand conditions. However, the ability to earn quasi-rents in the long run gives firms enjoying economies of scale the added advantage of more internal financial resources for growth as well as for research and development activities. Although there is considerable controversy concerning whether larger firms carry out proportionately more research and development than smaller firms, the outcome of the discussion is indecisive.[9] This is largely because industrial concentration is determined by economic and institutional factors that change very little over time and between countries. Consequently, it is unlikely to find enough variation in concentration within a single industry for a reliable test of a hypothesis concerning the impact of concentration on research and development. Comparisons among industries are not very reliable because of interindustry differences in the nature of R and D activities.

Market Structure and Product Definition

The concept of industrial concentration as a measure of monopoly power can be misleading without considering the question of product definition. A product can be defined in such a way that the "industry" is highly concentrated, yet there may be substantial competition from similar products of another industry. Joan Robinson suggests that a meaningful analysis of market structure must define an industry where the product is "bounded on all sides by a marked gap between itself and

[9] See Edwin Mansfield, *Technological Change* (New York: W. W. Norton, 1971), pp. 68–71.

its closest substitutes."[10] While this is easy to state in theory, the problems in delineating such an industry in the real world are overwhelming. Although the Department of Commerce has a standard industrial classification, individual firms may move from one industry classification to another on a yearly basis. Consequently, use of empirical data on industrial concentration as a measure of "market power" may be deceiving.

9-3 MARKET STRUCTURE AND MARKET PERFORMANCE

Economic theory has traditionally developed models of market performance that are tied to market structure. That is, given a stipulated market structure, firms would be expected to make price and output decisions in a characteristic way. We have already examined one such model, the market model, in which performance based on profit maximization depended on the structure of the industry. This model is also called the model of **perfect competition.** Another model describes the behavior of **monopolies,** industries consisting of a single firm. Both the models of perfect competition and of monopoly are ideal type models, since they are not intended to refer to market structures that actually exist. However, they are designed to improve our understanding of behavior in actual market structures. These include **duopolies** (two dominant firms) and **oligopolies** (a few dominant firms) as well as **monopolistically competitive** markets (many firms producing a slightly differentiated product). We shall consider each of these in turn.

Perfect Competition

The model of perfect competition presupposes an industry consisting of firms all producing an identical product.[11] Each firm is so small in relation to the total market output that its operations have a negligible impact on quantity supplied or on price. The individual firm can expand or contract output or leave the industry altogether with no significant impact on the market conditions. Of course, if all firms expand or contract output, market price will be affected. Firms are free to enter or leave the industry (there are no natural or institutional barriers), and all firms have perfect knowledge of industry conditions. All relevant parameters, such as profits being earned by other firms, the most efficient technological processes, market demand, and so on, are known and taken into account by all firms in the industry.

Given such a market structure, each firm is assumed to perform so as to maximize its individual profits. From this postulate of behavior we can derive certain characteristics of market performance in this model.

---

[10] Joan Robinson, *The Economics of Imperfect Competition,* 2d ed. (London: Macmillan and Co., Ltd., 1969), p. 17.

[11] The model can also be applied to a factor market. In this case, the factor services being offered must be identical.

In the short run firms will produce until the cost of the last unit, marginal cost, equals the product price, that is,

$$p = MC$$

provided marginal cost is increasing. Rising marginal cost is crucial to the model of perfect competition, since with decreasing or constant costs the profit-maximizing firm would expand indefinitely and eventually be large enough to influence market supply.

The short-run equilibrium position of the firm in perfect competition is shown in Figure 9–4. The rising portion of the marginal cost curve (above the average variable cost curve) can be viewed as the supply curve of the individual firm. Given the market price, quantity supplied can be found from the *MC* curve,

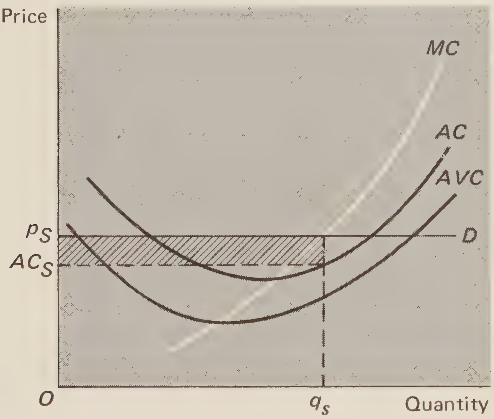


FIGURE 9–4
Perfectly Competitive Firm in Short-Run Equilibrium

provided *MC* is rising and is sufficient to cover variable unit costs. In a perfectly competitive market the supply curve is the horizontal sum of the supply curves of the individual firms.[12] The perfectly competitive model is the only case in which we can derive a supply curve independent of market demand.

In Figure 9–4 the firm is earning the quasi-rent indicated by the shaded area. The existence of this quasi-rent will attract new firms into the industry until quasi-rents disappear because of rising costs and falling product price. This zero profit equilibrium will occur at the point of minimum long-run average cost, as shown

[12] As industry output expands, increased demand for some factors of production specialized for that industry may cause these factor prices to rise, causing the *MC* curves for all firms to shift. Such shifts must be taken into account when aggregating the supply curves of the individual firms to produce a market supply curve. Economies and diseconomies external to firms but internal to the industry will have the same effect and should be handled similarly.

in Figure 9–5. The firm is operating in the most efficient plant size at minimum long-run and short-run average cost. This implies that

$$p_L = SMC = LMC = SAC = LAC$$

and all output is distributed in the form of payments to factors of production.[13] Although quasi-rents may still be earned by exceptional firms with locational or managerial advantages, the least efficient firms in the industry will operate at this point in the long run.[14]

It is important to keep in mind that the model of perfect competition is a theoretical construct designed to show the conditions under which Pareto efficiency would exist in a market system. The very notions of long-run equilibrium and perfect knowledge, for example, are at odds with the dynamic process

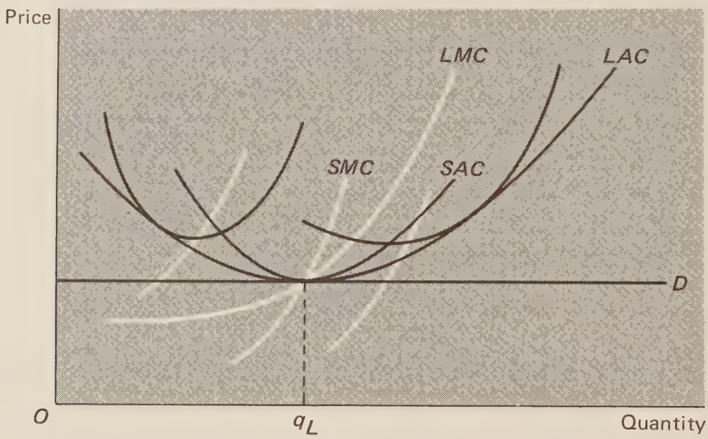


FIGURE 9–5
Perfectly Competitive Firm in Long-Run Equilibrium

that typifies an actual economy. Demand conditions, technology, factor prices and availabilities are constantly in flux, so that firms continually move in and out of industries. Furthermore, the absence of economies of scale makes the model inapplicable to most modern industries.

Yet the model may be applicable in a large number of individual cases, even though these do not represent the bulk of the value added in production. Recall that the 419,000 smallest companies in the United States economy account for

[13] In this case, paying factors their marginal value product exhausts the product regardless of the specification of the production function, that is, whether or not it is linear homogeneous.

[14] Such quasi-rents may be capitalized into factor payments in the long run. For instance, an exceptional manager may demand and receive a salary equal to the quasi-rent he generates if he threatens to leave the firm. In this case, in the long run all firms will be operating at the zero profit point where long-run average cost is minimized.

only 25 per cent of the total assets of United States manufacturing industry. This implies that the typical manufacturing industry is composed of a few large firms that establish price patterns and several hundred or thousand "fringe" enterprises that take prices as given and act as perfect competitors. This does not imply that market decisions in those industries are Pareto optimal, since the prices established by the large firms may depart from the Pareto criteria. Nevertheless, the model of perfect competition may be useful in analyzing the behavior of small firms in response to the prices established by few dominant firms.

Monopoly

The opposite extreme from perfect competition is a monopolistic market structure. A monopoly is an industry consisting of a single firm. Obviously, no industry consists of a single firm unless the product is so narrowly defined that there is extensive competition from close substitutes. The model of monopoly, like the model of perfect competition, is a hypothetical construct for analyzing how a firm would behave in isolation from any competition at all.

Any industry experiences some degree of competition from close substitutes, and this is taken into account in constructing the industry demand curve. In general, the greater the competition from substitutes the more elastic will be the market demand curve. For instance, the demand for tea is likely to be more elastic than the demand for all beverages taken together. If prices rise, consumers will leave the market for the substitute commodity. Lower prices will attract consumers from the substitutes. The analysis of monopoly takes the market demand curve as given and focuses on the behavior of the firm under those demand conditions.[15]

Once again we will assume that the objective of the firm is to maximize profits. Profit maximization occurs where any further increases in output will entail an addition to cost, marginal cost, in excess of the additional revenue, marginal revenue. For the monopolist, marginal revenue is less than price, since the operations of a monopolist have a significant impact on market supply. A monopolist can only increase output at the expense of falling prices. Thus the increment to revenue associated with selling an additional unit must take into account the price reduction *on all units sold,* which occurs when supply is increased. Thus,

$$MR = \frac{\Delta TR}{\Delta q} = p + q\left(\frac{\Delta p}{\Delta q}\right)$$

Since $\Delta p/\Delta q$ is negative, marginal revenue will always be less than price. The MR curve will always lie below the demand curve, as shown in Figure 9-6.

The profit-maximizing output, q_m, will occur where the MR curve intersects

[15] The model of the firm in monopolistic competition takes into account advertising and product variations as a means of attracting consumers from other industries.

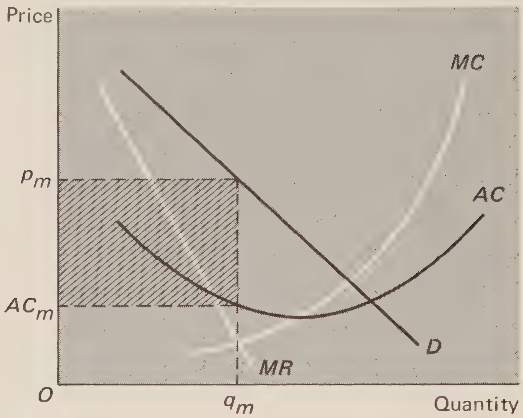
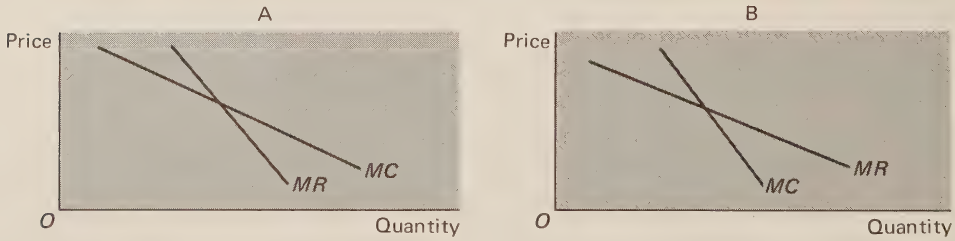


FIGURE 9–6
Profit Maximization in Monopoly (Short-Run)

the *MC* curve as long as *MR* intersects *MC* from above.[16] The price required to clear the market when q_m is produced is found on the demand curve at p_m. The difference between p_m and average cost at q_m is profit per unit. The shaded area represents total profits.[17]

The profit-maximizing equilibrium shown in Figure 9–6 must be viewed as a short-run situation. The existence of quasi-rents may attract more firms into the industry if p_m is above the entry-preventing price. Consequently, the monopoly must evaluate alternative strategies in the long run: Should we operate at lower profits and monopolize the industry, or would it be better to allow one or more additional firms to enter the industry and attempt to work out a market-sharing arrangement? The answer will depend on the amount of quasi-rents that can be earned if price is reduced to the entry-preventing level relative to the amount that would be earned if the market were shared by other firms. This, in turn, depends

[16] This will occur if *MR* is falling and *MC* rising. However, unlike the competitive case, a maximum profit position may also occur when *MC* is decreasing. In part A, $MR = MC$ produces a maximum profit. In part B, $MR = MC$ would produce minimum profits and the firm could expand indefinitely, increasing profits all the while.



[17] Monopoly profits are equivalent to rents or quasi-rents and must be distinguished from the return to capital established in the capital market. The payment to capital as a factor of production is included in the average cost curve.

on the optimal size of the firm in relation to the market as well as on the elasticity of demand. In addition, the firm must predict the behavior of its rivals should they be permitted to enter the industry. To evaluate such behavior we will examine some models of oligopolistic market structure. In any event, long-run equilibrium for a monopoly is consistent with short-run profit maximization only if p_m is an entry-preventing price. If it is not, market performance will depend on whether or not the firm decides to permit entry.

Monopoly and Perfect Competition Compared

Since a perfectly competitive market structure produces efficient resource allocation in production, it is often used as a standard to which market performance in other market structures is compared. Figure 9–7 shows the market equilibrium price and output for a monopoly, p_m and q_m, and for a perfectly competitive industry, p_c and q_c. The monopolist produces where MR equals MC. The perfectly competitive industry produces where price equals MC. MR is always less than price, and the monopoly output will always be less than the competitive output, assuming marginal cost is the same under both market structures. Since q_c is Pareto efficient, the monopoly underproduces from the point of view of efficient resource allocation.

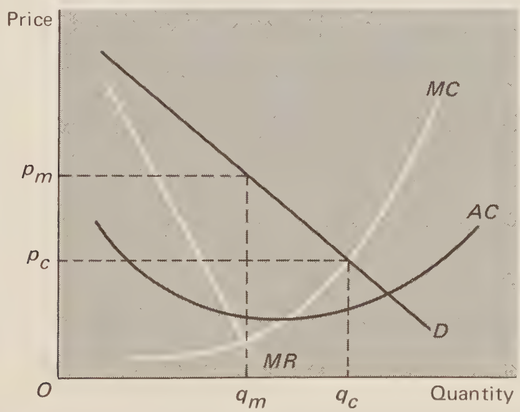


FIGURE 9–7
Monopoly and Perfect Competition Compared (Short-Run Equilibrium)

The extent of monopolistic underproduction depends upon the elasticity of market demand. In monopoly

$$MR = MC$$

while the efficient level of production occurs where

$$p = MC$$

But recall that

$$MR = p + q\left(\frac{\Delta p}{\Delta q}\right)$$

and elasticity of demand, E, is

$$E = -\frac{p}{q} \times \frac{\Delta q}{\Delta p}$$

Therefore

$$MR = p\left(1 - \frac{1}{E}\right)$$

so that for a monopoly,

$$MR = p_m\left(1 - \frac{1}{E}\right) = MC$$

or

$$p_m = \frac{MC}{1 - (1/E)} = MC\left(\frac{E-1}{E}\right)$$

The ratio between p_c, the competitive price, and p_m, the monopoly price, is $1 - (1/E)$, since

$$\frac{P_c}{P_m} = \frac{MC}{MC \div [1 - (1/E)]} = 1 - \frac{1}{E}$$

This implies that the more inelastic the demand the greater the difference between the competitive and the monopoly prices. The elasticity of market demand is often referred to as the "degree of monopoly," because it determines the extent to which monopolies can raise prices above the competitive level. If the availability of close substitutes makes market demand highly elastic, then monopolistic market performance may be quite similar to that of a competitive industry.

Notice that a monopolist will never operate where the elasticity of demand is less than 1, that is, where marginal revenue is less than zero. If marginal costs are positive, a monopolist will choose a point where marginal revenue is positive, so that demand elasticity must be greater than 1. This is not true for a competitive industry, which may operate in any range of the market demand curve.

Differences in Cost

Our comparison so far has assumed that a monopoly will have the same cost function as a perfectly competitive industry. From what we know about the

economic factors underlying differences in market structure, this is unlikely to be the case. If a firm is able to monopolize an industry because of economies of scale, then production by many small firms would push production costs up. In Figure 9-8 we have shown a situation in which marginal costs for a monopoly, MC_m, are lower than for a perfectly competitive market structure, MC_c. In this case, even though the monopoly sets MR equal to MC_m while the perfect competitors set price equal to MC_c, the monopoly price, p_m, is actually lower and the monopoly output, q_m, higher than for the competitive industry.

The Pareto criterion implicitly assumes that cost functions are independent of market structure. Clearly, however, if the antitrust division followed a policy of splitting up automobile manufacturers into units so small that they would have no influence on market supply, we would lose in technological efficiency much of what we would gain in allocative efficiency. An alternative would be to

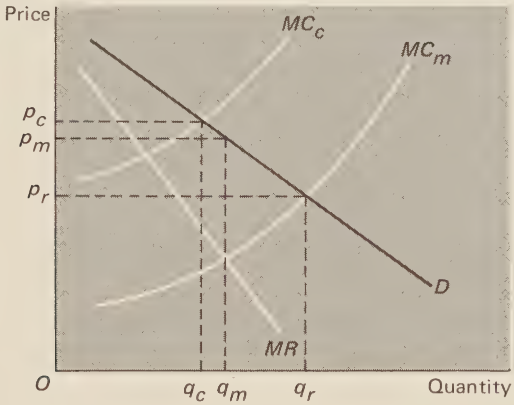


FIGURE 9-8
Monopoly and Perfect Competition with Different Cost Functions

allow the industry to develop the most efficient market structure from a technological point of view and then regulate it to produce where price equals marginal cost. In Figure 9-8 this occurs at q_r at a price of p_r. This philosophy is, of course, at the heart of the Lange system of decentralized socialism, in which firm managers are told to produce where price equals MC regardless of the size of the firm and regardless of the effect on profits.

In the absence of an underlying social philosophy that permits public control of industry, government policies toward business generally reflect recognition of a trade-off between technological and allocative efficiency in the presence of economies of scale. However, there is considerable controversy concerning the need for the corporate giants that exist in the American economy to exploit economies of scale. Many critics argue that for most American industries, pro-

duction could occur as efficiently in much smaller companies.[18] Furthermore, they claim that size in itself is not necessary to stimulate research and development and technological change, often attributed for the most part to firms holding monopoly power.[19] Others, like John Kenneth Galbraith, argue that only very large firms can afford to carry out the developmental activities upon which economic progress depends.[20]

Antitrust policy that is designed to reduce inefficiency in resource allocation must be realistic enough to recognize that perfectly competitive market structures are not technologically efficient in the presence of economies of scale. Consequently, the relevant question is whether or not breaking up corporate giants into a few more firms is desirable. That is, will more firms perform better than fewer firms when the industry is not perfectly competitive? To answer this question we turn to the theory of oligopoly in section 9–4.

Discriminating Monopoly

A monopolist cannot sell all he produces at a fixed price and so, if demand is fixed, he must reduce his price to increase sales. Because he must reduce the price on all units when the price of the last unit is lowered, marginal revenue, the additional revenue associated with the last unit sold, declines even faster than the product price. Suppose the monopolist were able to charge different prices to different users, so that customers who were willing to pay a higher price would pay that price but prices on additional units would be reduced to attract more buyers into the market. Since prices on the first units sold would not be reduced, marginal revenue would not decline as rapidly and sales and profits would be increased by more than in the case of the nondiscriminating monopolist.

Retail firms are well known for such discrimination in the form of reduced-price sales. Manufacturers often discriminate regionally or by changing the product slightly. Publishers distribute expensive, hardback editions of their books, then less expensive book club editions, and finally the least expensive paperbacks. The price differentials are much greater than justified by differences in cost. Service industries discriminate between categories of customers. Power rates are generally higher for domestic users than for commercial users, although these differences are not accounted for by differences in the marginal costs of the services.[21]

Two cases can be considered: First, a situation where a firm can segment its market into several categories of customers results in the establishment of differ-

---

[18] See George J. Stigler, "The Case Against Big Business," in Mansfield, *Monopoly Power and Economic Performance,* pp. 3–12.

[19] See Mansfield, *Technological Change,* pp. 68–71.

[20] John Kenneth Galbraith, *American Capitalism* (Boston: Houghton Mifflin, 1956), chapter 7.

[21] Price discrimination unjustified by differences in cost is technically illegal under the antitrust laws, but many firms are able to justify it in terms of cost differentials. In public utilities, for instance, overhead costs can be allocated in such a way as to demonstrate higher average total costs for domestic customers. Price differentials need not be justified by differences in *marginal* costs.

ent prices in those market segments. In this case, we are interested in determining what governs the differences in those prices. Second, in an extreme case of price discrimination the firm can charge a different price for each unit sold. This means that the firm can extract all the consumer surplus from the market in the form of profits. In this case, we can show how much of the good will be offered for sale.

Suppose an electric power company can discriminate between domestic and commercial users. If domestic users have fewer alternative sources of power than commercial users, or if households are less sensitive to the relative prices of electric power and other power sources (since expenditures on power are a smaller portion of a household's budget than of a commercial user's), the demand curve for domestic users will be less elastic than for commercial users.

If the monopolist is maximizing profit, he will set marginal revenue in each market equal to marginal cost of the output as a whole. If this were not true, the monopolist could increase profit by selling more in the market where marginal revenue is higher and less where marginal revenue is lower. Assuming production costs to be the same in each market, this implies

$$MR_D = MR_C = MC$$

where MR_D is the marginal revenue in the domestic market and MR_C is the marginal revenue in the commercial market.

Since marginal revenue is equal in both markets, prices will be different as long as demand elasticity is different. Recall that

$$MR = p\left(1 - \frac{1}{E}\right)$$

where E is elasticity of demand and p is the product price. Consequently, when marginal revenue is equalized in the two markets,

$$MR_D = p_D\left(1 - \frac{1}{E_D}\right) = p_C\left(1 - \frac{1}{E_C}\right) = MR_C$$

or

$$\frac{p_D}{p_C} = \frac{1 - (1/E_C)}{1 - (1/E_D)}$$

Since

$$E_D < E_C$$

then

$$p_D > p_C$$

that is, prices will be higher in the domestic market, where demand is less elastic.

The monopolist could increase his profits even more if he could further subdivide the market. Consider the extreme case of a monopolist who could

charge a different price for each unit sold. This **perfectly discriminating monopolist** could sell each unit for the maximum amount the consumer is willing to pay, so that his marginal revenue curve is the market demand curve. In Figure 9–9 the perfectly discriminating monopolist sells q_1 units at p_1, q_2 at p_2, q_3 at p_3, and so on, adding to his profits as long as the price paid for the additional unit exceeds marginal cost. Thus he will expand his level of sales as long as the demand curve lies above the marginal cost curve. Equilibrium occurs at (q_e, p_e) where price equals marginal cost. Total revenue is the area $OZEq_e$. This includes not only the revenue he would have gained as a nondiscriminator, Op_eEq_e, but also the consumer surplus, p_eZE.[22]

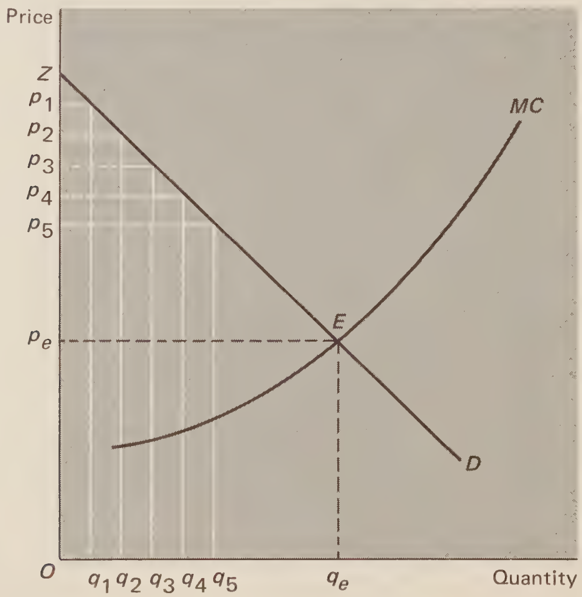


FIGURE 9–9
Perfectly Discriminating Monopoly

The quantity produced by the perfectly discriminating monopolist is Pareto efficient, because price equals marginal cost in equilibrium. However, since consumers are charged different prices, the distributive implications are different from perfect competition. The effect is the same as if the monopolist had been subsidized to force him to produce where price equals marginal cost, with the subsidy being raised by a tax on consumers' incomes. Although price discrimination will not affect consumers' demand for the product (except through the

[22] As a nondiscriminator, the monopolist would not produce at q_e but at the output for which marginal cost equals marginal revenue, so that the consumer surplus overstates the gains associated with discriminatory pricing. Consumer surplus is the difference between the profits of a perfectly discriminating monopolist and a marginal cost pricer.

income effect), it will have a different distributive impact from a uniform tax on income, hurting most those people who are willing to pay relatively high prices for the commodity.

Where the monopolist can discriminate among several categories of buyers, but where discrimination is not perfect, output will be greater than the normal (nondiscriminating) monopoly output and profits greater than monopoly profits, but below the Pareto efficient output. The greater the ability to discriminate, the greater the monopoly profits and the closer the output to the Pareto efficient level.

9-4 DUOPOLY AND OLIGOPOLY

It is often supposed that because a perfectly competitive market structure is more conducive to allocative efficiency than monopoly, the more firms there are in an industry (consistent with technological efficiency) the better that industry will perform in terms of the efficiency criteria. Thus, if technological considerations would permit smaller firms to produce under the same efficient cost functions as the corporate giants, then the larger firms should be broken up in the interests of allocative efficiency.

The analogy is false, however, since the perfectly competitive model presupposes that all firms are so small as to have no impact on market supply. We need a model of market behavior in which there are several firms, each large enough to influence market supply. Such a model could answer the question: Is oligopoly better than monopoly?

The salient feature of all oligopoly models is that the sellers are independent but their actions are interdependent.[23] For any firm, i, the effect of a change in supply by another firm, j, on its profits, π_i, is not negligible, that is,

$$\frac{\Delta \pi_i}{\Delta q_j} \neq 0$$

In perfect competition, of course,

$$\frac{\Delta \pi_i}{\Delta q_j} \approx 0$$

where $\approx$ means approximately equal.

Not only must a firm in an oligopolistic market evaluate its own profit-maximizing position but it also must figure out what its rivals are going to do. Because oligopolistic market performance depends upon a subjective evaluation of the behavior of each participant firm by each participant firm, there are many possible models that may be relevant in different situations. There may be attempts at collusion where demand is inelastic, because the impact on price of any individual's action can have serious consequences for the other participant

[23] If the sellers are not independent, they can be treated as a monopoly.

firms. On the other hand, firms may take a passive role, merely reacting to the behavior of their rivals. Although objective economic considerations (demand elasticity, control over important resources, ability to differentiate the product) may provide insight into the type of behavior that would characterize a particular industry, in general there is no a priori way to apply particular behavior models to individual industries in the real world. Each case should be evaluated separately.

There are almost as many different models of oligopolistic behavior as there are economists interested in the subject. You may even wish to develop your own. In this section we present several historical models, which are interesting more for their methodological implications than for empirical validity, as well as some modern ones that purport to describe actual behavior.

The Cournot Model

One of the earliest attempts to describe oligopolistic market behavior was made by Augustin Cournot. Cournot's model provides the methodological framework for a rather general statement of the oligopoly problem, and a number of other models have been based on his assumptions. To simplify matters, Cournot analyzes a duopoly, that is, a market consisting of two sellers. It is easily generalized for the case of three or more sellers.

The basic characteristic of the Cournot model is that each participant firm assumes a passive role, taking its rivals' actions as given. Each firm attempts to maximize its individual profits subject to the (passively accepted) actions of the other participant firms. There is no attempt by the firms to increase their profits by collusion or to influence each other in any way. The product is assumed to be homogeneous, so that a firm cannot attain an advantage by product differentiation.[24]

Consider the case of an industry composed of two sellers. Total market output is the sum of the outputs of the two firms, that is,

$$q = q_1 + q_2$$

The price, p, at which the market output can be sold can be determined from the demand function

$$p = f(q) = f(q_1 + q_2)$$

This price will depend on the output of both participants.

For any firm, total revenue, pq_i, will depend not only on his own output but also on that of his rival, since

$$TR_1 = pq_1 = f(q_1 + q_2)q_1$$
$$TR_2 = pq_2 = f(q_1 + q_2)q_2$$

[24] As we will see, this assumption is not necessary to the analysis.

Similarly, the profit functions of the two firms are interdependent,[25] since

$$\pi_1 = TR_1 - TC_1 = f(q_1 + q_2)q_1 - TC_1(q_1)$$
$$\pi_2 = TR_2 - TC_2 = f(q_1 + q_2)q_2 - TC_2(q_2)$$

Each firm maximizes profit by setting marginal revenue equal to marginal cost. The profit-maximization equation for each firm depends on the output of both firms, that is,

$$MR_1(q_1, q_2) - MC_1(q_1) = 0$$
$$MR_2(q_1, q_2) - MC_2(q_2) = 0$$

These equations are called **reaction functions.** They can be solved simultaneously to determine q_1 and q_2.

Suppose, for example,

$$MR_1 = 100 - q_1 - 0.5q_2$$
$$MR_2 = 100 - q_2 - 0.5q_1$$
$$MC_1 = 5$$
$$MC_2 = q_2$$

Then the reaction function ϕ_i for each firm would be

$$\phi_1 = 100 - q_1 - 0.5q_2 - 5 = 95 - q_1 - 0.5q_2 = 0$$
$$\phi_2 = 100 - q_2 - 0.5q_1 - q_2 = 100 - 0.5q_1 - 2q_2 = 0$$

When these are solved simultaneously, the output for each firm is found to be

$$q_1 = 80$$
$$q_2 = 30$$

The market price can be found by substituting into the demand function. Suppose the demand function is

$$p = 100 - 0.5(q_1 + q_2)$$

Then

$$p = 100 - 0.5(80 + 30) = 100 - 55 = 45$$

The solution can also be found graphically, as shown in Figure 9–10. When the reaction functions are graphed in (q_1, q_2) space, duopoly equilibrium is found at their intersection.

The analysis can be applied to more than two firms. If there are n firms, there will be n reaction functions. Since the market price will depend on the outputs of the n firms, that is,

$$p = f(q_1 + q_2 + \cdots + q_n)$$

[25] Cournot assumed that profit functions are interdependent because of the relation between the firms' outputs in the revenue function. In an industry with only a few firms cost functions may also be interdependent because of pecuniary and technological externalities.

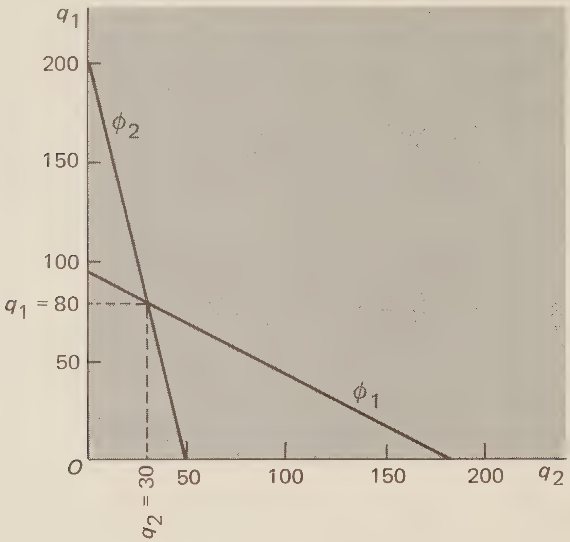


FIGURE 9–10
Duopolistic Equilibrium in a Cournot Model

the variables $q_1, q_2, \ldots, q_n$ will appear in the *MR* function of each firm and consequently in each firm's reaction function. Thus there will be *n* variables and *n* equations, so that a solution can, in general, be found.[26]

In the Cournot model, as the number of firms becomes very large, the output of any one of them has a negligible impact on market supply and the solution approaches the competitive case, in which each firm can sell all it can produce at the market price and consequently sets price equal to marginal cost. If all firms are producing under conditions of constant cost, then as more firms enter the industry, market price will fall, output will increase, and the equilibrium approaches competitive equilibrium.[27] Thus, on the basis of this model, more firms are preferred to fewer. However, when cost is taken into account, this is not necessarily the case. If there are economies of scale, then increasing the number of firms may sacrifice technological for allocative efficiency.

The Stackelberg Model

The Cournot model emphasized the importance of the *number* of sellers for oligopolistic equilibrium. The Stackelberg model, while methodologically similar to the Cournot model, emphasizes differences in the *behavior patterns* of participant firms. The Stackelberg model distinguishes between **leaders** and **followers.** A follower adjusts his output level to maximize profits, assuming that the rival is

[26] This is, of course, subject to the possibility that *n* equations cannot always determine *n* variables, as mentioned in connection with the theory of general equilibrium.

[27] This is also true when marginal costs are increasing and all firms have identical cost functions.

the leader and makes his output decisions independently. The leader assumes the rival will act as a follower and maximizes his profit subject to his rival's reaction function.

Suppose, as before,

$$\begin{aligned} MR_1 &= 100 - q_1 - 0.5q_2 \\ MR_2 &= 100 - q_2 - 0.5q_1 \\ MC_1 &= 5 \\ MC_2 &= q_2 \end{aligned}$$

Assume also that firm 1 is a leader and firm 2 is a follower. Then, for firm 2, the reaction function is

$$\phi_2 = 100 - 0.5q_1 - 2q_2 = 0$$

or

$$q_2 = 50 - 0.25q_1$$

Firm 1 substitutes this value for q_2 into its profit function and maximizes profits, which are now a function of q_1 only. Assuming that the market demand function is

$$p = 100 - 0.5(q_1 + q_2)$$

then

$$\begin{aligned} \pi_1 &= pq_1 - TC_1(q_1) \\ &= [100 - 0.5(q_1 + q_2)]q_1 - 5q_1 \\ &= \{100 - 0.5[q_1 + (50 - 0.25q_1)]\}q_1 - 5q_1 \\ &= 70q_1 - 0.375q_1^2 \end{aligned}$$

By applying simple calculus it can be shown that

$$\frac{\Delta\pi_1}{\Delta q_1} = MR_1 - MC_1 = 70 - 0.75q_1 = 0$$

or

$$q_1 = 93\tfrac{1}{3}$$

Substituting into ϕ_2,

$$q_2 = 50 - 0.25q_1 = 26\tfrac{2}{3}$$

Market price can be found from the demand function

$$\begin{aligned} p &= 100 - 0.5(q_1 + q_2) \\ &= 100 - 0.5(93\tfrac{1}{3} + 26\tfrac{2}{3}) \\ &= 40 \end{aligned}$$

In this case, leadership behavior produced a greater output and a lower price than the Cournot case, given the same market demand and cost functions for the individual firms. This is because the leader happened to be the low cost producer,

and in the leadership model the distribution of production was technologically more efficient than in the Cournot case.

The actual market equilibrium in any case—given the market demand function and the cost functions of the individual firms—will depend on which firms view themselves as leaders and which as followers. If all firms are followers, we have the Cournot case. If more than one firm decides to be a leader, however, an equilibrium may never be attained. This is because each leader assumes the other's behavior is determined by his reaction function, but it really is not. Thus the expectations of the leaders regarding their rivals' actions may not be realized, and they will readjust output. If the firms continue to view themselves as leaders and are constantly readjusting output, the market is said to be in **Stackelberg disequilibrium.**

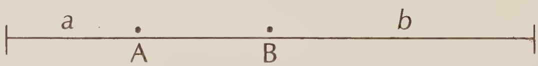
In this model the entry of additional firms may or may not have a destabilizing effect on the market, depending upon the type of behavior followed by the firm. Thus, in contrast to the Cournot case, fewer firms may perform better than more firms when variations in market roles of the participant firms are taken into account.

The Hotelling Model of Spatial Competition

We have assumed so far that the only way in which firms can improve their profits is by varying quantity supplied. In fact, given highly inelastic demand in some industries, firms may hesitate to vary market supply in order to improve their profits. Furthermore, collusive agreements may prevent them from altering selling price, so that the quantity they can sell at that price is fixed. In such industries firms may try to increase their profits by changing the product to attract new customers. One form of product variation is change in location. Hotelling developed a model of spatial competition, which examines optimal location patterns when quantity demanded is fixed.[28] The model has implications for all kinds of product competition.

Consider two firms A and B producing the same commodity. The total market demand for the commodity is completely inelastic, that is, there is a fixed amount purchased regardless of price. However, because of positive transportation costs, consumers will purchase from the firm that is most conveniently located.

Suppose the consumers are uniformly distributed along a line and the two firms A and B are located at distances *a* and *b* respectively from each end of the line:



Assume also that marginal costs are zero, so that the goal of each firm is to get as many consumers as possible. In this case, A will attract all the consumers in

[28] Harold Hotelling, "Stability in Competition," *Economic Journal, XXXIX* (March 1929), pp. 41–57.

a plus those in one-half the distance between A and B. B will attract all the consumers in *b* plus those in one-half the distance between A and B. Clearly, it will be in the interest of each firm to have as much of the line as possible behind it. A could increase its customers by moving next to B:

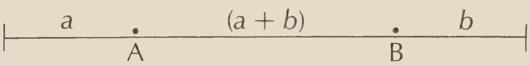
since now A attracts all the consumers in *a*. But B can do better by moving to the left of A:

The leftward move will continue, with A jumping over B, B over A, and so on, until they end up exactly in the middle, side by side:

They then have equal shares of the market, so that

$$a = b$$

From the viewpoint of the consumers the optimal location would be at the quartiles, since this location minimizes overall transportation costs:



The optimal location for the firms, however, is actually the least convenient for the consumers. Transportation costs are maximized when the two firms are located in the center.

Hotelling uses his model to explain the similarity of political parties in the United States:

> The competition for votes between the Republican and Democratic parties does not lead to a clear drawing of issues, an adoption of two strongly contrasted positions between which the voter may choose. Instead, each party strives to make its platform as much like the other's as possible. Any radical departure would lose many votes, even though it might lead to a stronger commendation of the party by some who would vote for it. Each candidate "pussy foots," replies ambiguously to questions, refuses to take a definite stand in any controversy for fear of losing votes.[29]

[29] Ibid., p. 55.

Although he was describing the political scene in 1928, he could well have been talking about contemporary politics. Even those candidates who take extreme positions at first generally back down when they find they are "serious contenders." Although this is sometimes attributed to the "moderation of political maturity," it may in fact be associated with the practicality of getting more votes than one's opponents. Candidates like Barry Goldwater, who refused to compromise extreme positions, are doomed to defeat. The convention that nominated George McGovern for the Presidency warned him that he would be a "Democratic Goldwater" unless he moderated his position on United States foreign policy, bringing it closer to that of President Nixon, the Republican incumbent.

The analysis has implications for product design. Year after year, competing car manufacturers come out with new models that are strikingly similar to each other. Deviants from the pattern for that year experience disastrous losses.

If demand is elastic, so that consumers actually leave the market if transportation costs are too high or if they dislike the product design, there will be an opposing tendency to move away from the center.[30] In the spatial example, A and B may move away from each other if customers at the far ends of the line are leaving the market because of high transportation costs. Thus the elasticity of demand provides a check to the move toward the center.[31]

It can be shown that the tendency for movement toward the center is still present as more firms enter the industry. A new automobile manufacturer would undoubtedly strive to make his product as much like the Ford and General Motors lines as possible. Thus, increasing the number of firms in an industry may not increase the variety of products available within an industry, despite advertising claims by manufacturers that their products are truly different and superior. Consumers like to believe that their tastes are highly individualistic, even though in the end most products they consume are pretty much like those consumed by their neighbors.

Kinked Demand Curve

The model of the kinked demand curve, originally developed by Paul Sweezy, is an attempt to explain why prices in oligopolistic markets tend to be so stable, despite many changes in demand or costs.[32] Why, in fact, do oligopolists tend to use product and locational variation much more than variation in quantity supplied? We suggested that this type of behavior occurs most frequently in industries characterized by inelastic market demand. Sweezy shows that for an individual oligopolistic firm, demand may be even less elastic than overall market demand.

---

[30] This point is made by Arthur Smithies, "Optimum Location in Spatial Competition," *Journal of Political Economy, XLIX* (June 1941), pp. 423–439.

[31] In the political example, of course, demand is completely inelastic, short of a revolution that would overthrow the presidential form of government.

[32] Paul M. Sweezy, "Demand Under Conditions of Oligopoly," *Journal of Political Economy, XLVII* (August 1939), pp. 568–573.

The Cournot model assumed that all market participants know each other's reaction functions. Suppose firms do not know how their rivals will behave. In order to avoid disastrous consequences associated with its own actions each firm must assume that its rivals will respond in the least favorable way. If the firm increases prices, the worst thing the rivals can do is to maintain their prices at a lower level. If the firm reduces prices, the least favorable reaction of the rivals is to reduce their prices too.

Under the circumstances, the oligopolistic firm must view its demand curve as similar to the one shown in Figure 9–11. Market price is established at p_s.

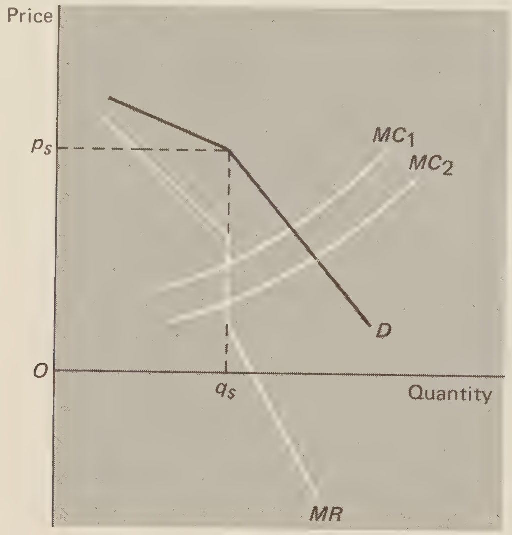


FIGURE 9–11
A Kinked Demand Curve

If the firm increases price and the rivals maintain prices at p_s, the firm will experience a greater reduction in quantity demanded than if the rivals had followed the price increase, and the demand curve will be relatively flat. On the other hand, if the firm decreases price and the others follow suit, there will not be as great an increase in quantity demanded as if firms maintained their prices, and the demand curve will be much steeper. Consequently, the demand curve is "kinked" at the established price, p_s.

The kink in the demand curve produces a gap in the marginal revenue curve at q_s. That is, marginal revenue associated with the demand curve to the right of q_s is substantially lower than to the left of it.[33] As shown in Figure 9–11, shifts in the marginal cost curve can occur without affecting output or price.

Although changes in cost conditions that affect individual firms may not affect

[33] The size of the gap will depend on the difference between the slopes of the upper and lower portions of the demand curve.

market performance in such an industry, industrywide cost changes may provide a signal for all firms to raise prices simultaneously. The establishment of safety requirements in the tire industry resulted in an across-the-board price increase despite the fact that the dominant firms were already meeting the safety requirements before they were made obligatory. Meeting safety standards occasioned little or no increase in cost yet provided a signal for an industrywide price increase. Because of rising labor costs for individual firms, tire prices would have been raised earlier if the firms had not been afraid their rivals would maintain prices at the previous level.

Tax increases also provide a signal for price increases. A tax on profits should have no effect on market performance, since it affects neither marginal revenue nor marginal cost. If a profit-maximizing position is established, the imposition of a tax on a per cent of those profits will not change the output at which profits are maximized. Nevertheless, Krzyzaniak and Musgrave found that the federal corporate profits tax in United States manufacturing industry has been shifted by more than 100 per cent in the short run. For every $1.00 increase in corporate tax liability, pretax profits increased by $1.34.[34] Presumably, market prices had been below the profit-maximizing level but firms were afraid to raise prices because of perceived kinks in their individual demand curves.

When unions negotiate wage increases on an industrywide basis, and when raw materials prices increase, firms may increase product price by more than if the market were a monopoly or if there were overt collusion. Many observers have attributed the phenomenon of cost-push inflation, which has characterized the American economy for the past twenty years, to the type of oligopolistic market behavior described in the kinked demand curve model.[35] Under these circumstances, tax increases which are designed to reduce the inflationary pressure of demand may actually contribute to raising prices.

Market Shares

In some industries overt collusion takes place. In others, for historical reasons, firms have tacitly accepted fixed market shares. Suppose there are three firms in the industry. One firm produces one half the total market output, while the other two each produce one fourth. The dominant firm may assume the role of the leader, knowing the reaction function of the other two firms:

$$q_2 + q_3 = \tfrac{1}{2}(q_1 + q_2 + q_3) = \tfrac{1}{2}\, q_1 + \tfrac{1}{2}(q_2 + q_3) = q_1$$

If all firms produced under identical constant cost conditions, the maximum profit position for the leader would also maximize profits for the rivals. Furthermore, the market equilibrium would be identical to that of a monopoly. Marginal cost is the same for all firms and is independent of firm size. Thus the maximum

[34] M. Krzyzaniak and R. Musgrave, *The Shifting of the Corporation Income Tax* (Baltimore: Johns Hopkins Press, 1963).

[35] See Nancy S. Barrett et al., *Prices and Wages in U.S. Manufacturing* (Lexington: D.C. Heath, 1973).

profit position for the individual firm will be the profit-maximizing position for the industry, and profit shares will be proportional to market shares.

Unless costs are constant and identical, however, firms in a market-sharing arrangement will generally charge higher prices and produce less than a pure monopoly. If costs are decreasing, for instance, it would be better for all production to take place in a single plant. If costs are increasing, production should be distributed so as to equalize marginal costs of the participant firms. If cost functions were identical, this would imply an equal distribution of production among the firms. Market sharing may, however, be conducive to product competition as firms attempt to increase their shares by nonprice competition. This may result in real product improvements from the viewpoint of the consumer. On the other hand, it may produce excessive competitive advertising and increased costs, which are pyramided into higher product prices.

From the viewpoint of the consumer, a profit-sharing arrangement would be preferable to an output-sharing arrangement, since in the former the group would attempt to maximize overall industry profits and hence organize production with regard to technological efficiency. In a market-sharing arrangement, the least efficient firms may actually receive a disproportionate share of the market for historical reasons. In any event, performance in a market-sharing model will never be better than in pure monopoly (a monopolist can always increase the number of plants under conditions of increasing costs), since market sharing does not always conform to the least cost distribution of production.

Behavioral Models of the Firm

Recently economists have questioned the assumption that the primary objective of large corporations is profit maximization. Firm managers are rarely the owners, and they may be concerned with maximizing their own incomes and power rather than company profits. The growing bureaucratization of industry suggests that the concept of a unified set of company objectives may also be obsolete.

Behavioral models of the firm examine the process by which decisions are made within firms and attempt to determine the objectives, if any, that management is trying to achieve. On the basis of such analysis a model of market performance can be developed. Some models emphasize the implications for market performance of changes in the objective function.[36] Others concentrate on developing empirical models of the decision process of individual firms.[37]

The Sales-Maximization Model

A model of the first type was developed by William Baumol. He argues that in cases where managers of a firm are not the owners, they are likely to be more

[36] See William J. Baumol, *Business Behavior, Value and Growth,* rev. ed. (New York: Harcourt, Brace and World, 1967), chapter 6. Models of this type are sometimes referred to as *managerial models* of the firm.

[37] R. M. Cyert and J. G. March, *A Behavioral Theory of the Firm* (Englewood Cliffs, N.J.: Prentice-Hall, 1963).

interested in maximizing the volume of sales than in maximizing profits. Not only are managers interested in a large power base, which is closely related to the scale of operations, but it is important to have a high volume of sales to maintain or increase the firm's share of the market.[38]

The firm maximizing total sales will produce in the unit elastic range of its demand curve, that is, where marginal revenue equals zero. Figure 9–12 compares the sales-maximizing equilibrium (p_s, q_s) with the profit-maximizing equilibrium (p_m, q_m). Notice that sales maximization always results in a lower price and correspondingly higher output than profit maximization. In the case shown in Figure 9–12, sales maximization results in performance very close to the competitive solution (p_c, q_c).[39]

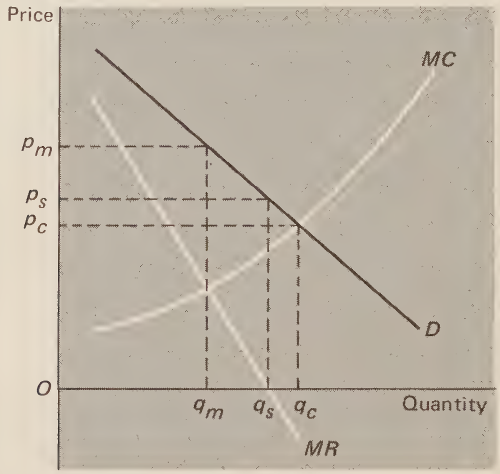


FIGURE 9–12
Sales Maximization Versus Profit Maximization

The sales-maximization model can be applied to oligopoly as well as to monopoly. In cases where the oligopolist was assumed to maximize profits subject to the actions of rivals, we could assume the firm maximizes sales instead. In the Cournot model, for instance, the firm's reaction function could be derived under the assumption that it maximizes sales rather than profits.

Notice that in the sales-maximization model, market performance is independent of costs. This means that in markets where firms operate under widely varying cost conditions, sales maximization will result in a considerably different distribution of production than profit maximization.

[38] For a discussion of some other reasons for selecting sales maximization as an objective see Baumol, *Business Behavior, Value and Growth,* chapter 6.

[39] If the *MC* curve intersects the demand curve in the unit elastic range, sales maximization will result in marginal cost pricing and allocative efficiency.

Markup (Administered) Pricing

Some studies of oligopolistic pricing practices suggest that many firms follow a markup approach. That is, they set price as a markup over average cost, where average cost is reckoned as some normal unit cost at a standard level of output.[40] Price is set according to the following formula:

$$p = (1 + \mu)(AC_N)$$

where μ is the markup factor and AC_N is the normal average cost at the standard level of output. Prices will change in proportion to changing factor prices or factor productivity (including technological change). A wage increase, for instance, will be fully passed on to consumers in a proportionate price increase. If taxes are included in the calculation of average cost, then tax changes will also be pyramided into higher prices. However, prices will not change in response to changes in the level of demand or changes in the scale of output.

The actual markup for any firm will depend upon barriers to entry for the industry and the degree of monopoly, reflected in the elasticity of demand. Thus changes in demand elasticity that are perceived by the firm will affect prices. If demand becomes more inelastic (or is perceived to become more inelastic), firms will increase their markup and prices will rise in the absence of an increase in costs. If costs are rising when demand elasticity is falling, price increases will be greater than the increase in costs. Falling demand elasticity can be interpreted as an increase in the market power or monopoly power of the firm. Thus demand conditions may affect pricing in this model if they affect the degree of monopoly, even though changes in the general level of demand are presumed to have no impact.

Although markup, or administered, pricing seems to be a rather common practice among oligopolistic firms,[41] many economists view it as a rule of thumb that will result in the maximization of some objective—profits, sales, or some other target variable—over the planning period. In other words, rules of thumb will only persist as decision criteria to the extent that they are consistent with the firm's ultimate objectives.[42]

Other economists argue that rules of thumb such as markup pricing do not necessarily result in optimal decisions with respect to some implicit objective but represent a more or less safe course of action for the firm in the face of uncertainty and imperfect information.[43] Since oligopolists (relative to participants

---

[40] For a discussion of how the "normal" cost and "standard" output are determined, see Otto Eckstein and Gary Fromm, "The Price Equation," *American Economic Review, LVIII* (December 1968), pp. 1159–1183.

[41] See, for example, the interview studies in A. D. N. Kaplan, J. B. Dirlam, and R. F. Lanzilotti, *Pricing in Big Business—A Case Approach* (Washington, D.C.: Brookings Institution, 1958).

[42] See William J. Baumol and Richard E. Quandt, "Rules of Thumb and Optimally Imperfect Decisions," *American Economic Review, LIV* (March 1964), pp. 23–46.

[43] See H. A. Simon, "Theories of Decision-Making in Economics," *American Economic Review, XLIX* (June 1959), pp. 253–283.

in other market structures) are most likely to be faced with uncertainty because of the inability to predict rivals' reactions to changing demand and supply conditions, markup pricing is most likely to characterize oligopolistic market structures. In this view, oligopolistic market behavior seeks satisfaction rather than maximization. That is, businessmen have criteria for satisfactory performance and are willing to adopt rules of thumb by which these criteria will be met even if such rules of thumb do not maximize their target variables.

Although studies have indicated that in some cases markup pricing does produce decisions similar to those that would have been reached using a profit-maximization criterion, the empirical evidence is not conclusive.[44] However, the markup pricing model is consistent with the observed tendency of product prices to fluctuate with factor prices in oligopolistic markets, producing the phenomenon of cost-push inflation. In fact, this model is often offered as an alternative to the kinked demand curve model. Both attempt to describe how oligopolies will behave in the face of uncertainty. And both conclude that changes in costs will have a greater effect on prices in oligopolistic markets than in monopolistic or perfectly competitive markets.

More Versus Fewer: An Evaluation

We began this section by pointing out that antitrust policy in the United States has traditionally been based on the assumption that more firms in an industry are better than fewer unless economies of scale are so great that technological efficiency would be compromised. Most studies related to the advisability of breaking up large firms have focused on the question of the optimal size of the firm with respect to costs. However, we have seen that certain models of price behavior suggest that monopoly or collusive agreements might result in lower and/or more stable prices and correspondingly higher outputs than noncollusive oligopoly. The Cournot model, in which firms passively react to each other's behavior, is the only one that produces the unequivocal result that prices fall as more firms enter the industry. And this result can only be achieved under the assumption of constant costs of production.

In the Stackelberg model the entry of new firms may produce price instability if more than one firm attempts to be a leader. Market price may be above or below the monopoly price, depending on the cost functions and distribution of production among leaders and followers.

The kinked demand curve model would produce short-run price stability, but this type of behavior is conducive to cost-push inflation. Administered pricing also translates rising costs into higher prices. Monopoly prices would rise more gradually in response to rising costs and would not promote the type of inflationary psychology that occurs when cost increases are matched by the equal and sometimes greater price increases, which are forthcoming in oligopo-

[44] For an example of an empirical study of this issue see Eckstein and Fromm, "The Price Equation."

lies following the kinked demand curve pattern of market performance or markup pricing.

Finally, we saw that monopoly is generally preferable to market-sharing arrangements. However, firms may attempt to increase their market shares by making real product improvements that benefit the consumer. On the other hand, nonprice competition may take the form of excessive advertising with little or no product improvement, which raises production costs and market prices. Although there may be an advantage to the availability of a large variety of products, presumably a monopoly will also resort to product variation to increase profits. Automobile manufacturers, cigarette companies, producers of soap, toothpaste, and so on, all offer multiple product lines. Product differences in the automobile industry, for instance, are greater within individual firms than between firms.

Comparisons are often made between the variety of consumer products in a capitalist economy and the rather uniform product lines that characterize centrally planned economies. It is likely that this difference is due to the absence of a profit motive for the individual firm under socialism rather than to the monopolization of industry per se. On the other hand, unnecessary (noninformational) advertising and excessive diversification of products is likely to occur in the capitalist oligopoly, where firms are attempting to increase market shares. Monopoly would find such activities unprofitable.

Since theoretical evidence indicates that monopoly or overt collusion is in many cases preferable to oligopoly, it is not surprising that many European countries actually encourage rationalization of industry (increasing the degree of concentration) and cartel arrangements (open collusion). In the United States, however, there is an overriding fear of the great concentrations of economic power that result from the growth of firms. The problem is more serious in the United States because markets are larger and the potential size of monopolistic companies is greater. In addition, government does not have the control over industry that it has in many European countries, where large firms are often semipublic enterprises. Thus antitrust measures cannot be based on economic considerations alone. Industrial structure has serious social and political consequences as well as implications for economic performance.

9-5 MONOPOLISTIC COMPETITION

The model of monopolistic competition was developed by Edward H. Chamberlin to explain the effect of product differentiation on competitive market performance.[45] He assumed there are many small firms in the market, each producing an amount that has a negligible effect on market supply, yet each firm produces a

---

[45] Edward H. Chamberlin, *The Theory of Monopolistic Competition*, 8th ed. (Cambridge, Mass.: Harvard University Press, 1962).

slightly differentiated product.[46] Chamberlin defines a differentiated product as one for which

> . . . any significant basis exists for distinguishing the goods (or services) of one seller from those of another. Such basis may be real or fancied, so long as it is of any importance whatever to buyers, and leads to a preference of one variety of the product over another. . . . Differentiation can be based upon . . . pecularities of the package or container . . . or singularity in quality, design, color, and style.[47]

Chamberlin argued that the pervasiveness of advertising in the American economy has resulted in at least a perception of product differentiation in most industries. Brand loyality will induce some people to purchase products that are more expensive than products with exactly the same objective characteristics. A housewife may buy a well-known brand of laundry detergent for 50 cents a pound when an equally effective store brand is available for 32 cents. The housewife may justify her purchase on the grounds that the store brand is an inferior product, but rarely can she give you evidence that this is so. She never tried the store brand, but she saw a television commercial (which she knows was written by the makers of the well-known detergent) that said it gets clothes whiter. More than likely, the housewife has no idea which is the better product, but she pays more for the well-known brand because she feels happier with it.

Because some consumers will continue to purchase its products when a firm raises its price above that of its competitors, the firm's demand curve will be negatively sloped. Thus, even though the firm is small in relation to the total market, it possesses a degree of monopoly that is measured by the elasticity of the demand curve. The more differentiated the product from close competitors, or the greater the brand loyalty of its customers, the more inelastic will be the demand curve of the firm and the greater the degree of monopoly.

In the short run the monopolistic competitor will maximize profits where marginal revenue equals marginal cost. But unlike the monopoly or oligopoly for which economies of scale exist, the monopolistically competitive firm cannot prevent entry into the industry when it is earning quasi-rents. Consequently, firms will enter as long as quasi-rents are positive. Long-run equilibrium (p_m, q_m) occurs where price equals average cost and quasi-rents are zero, as shown in Figure 9–13.

Comparison with the Competitive Model

In the competitive model the firm's demand curve is horizontal, and long-run equilibrium occurs at minimum long-run average cost q_c in Figure 9–13. Notice

[46] The assumption of many small firms was an attempt to isolate the effect of product differentiation from that of economic barriers to entry due to economies of scale. Product differentiation and product competition can also occur in oligopolies where the participants are afraid to risk price competition, particularly where overall industry demand is inelastic. The tobacco industry is a case in point.

[47] Chamberlin, *Theory of Monopolistic Competition,* p. 56.

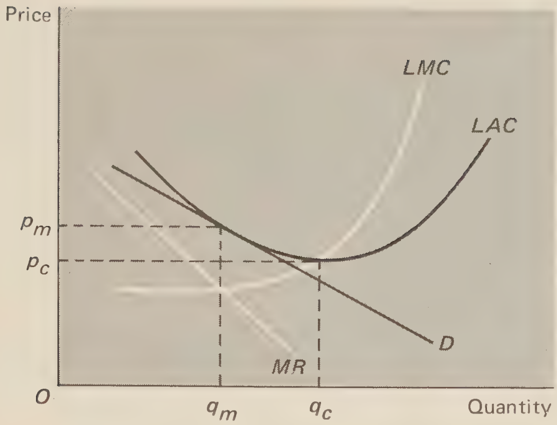


FIGURE 9–13
Long-Run Equilibrium for a Monopolistically Competitive Firm

that the long-run equilibrium price in monopolistic competition, p_m, is higher than the long-run competitive equilibrium price, p_c. Furthermore, at q_m price is above marginal cost, so that resource allocation is inefficient. In addition, at q_m the firm is producing in the decreasing range of the *AC* curve, at a higher unit cost than at q_c. This implies that the same amount could be produced more cheaply by fewer firms. Chamberlin argues that the industry in the long run is producing under conditions of excess capacity.

He suggests that from the viewpoint of efficiency monopolistic competition may be wasteful because it requires more firms to produce the same output at higher costs than in perfect competition. On the other hand, this "waste" may be justified on the basis that more variety represents a gain in utility to the consuming public. Higher prices under monopolistic competition can be viewed as the cost of product differentiation. If consumers are willing to pay the higher amount, this means the social benefit is at least equal to the cost.

Selling Costs and Advertising

Any evaluation of the welfare implications of Chamberlin's model must take into account the effect of advertising outlays on selling cost as well as the implication of advertising for consumer well-being. Firms can increase their profits by advertising if advertising reduces the elasticity of their demand curve or shifts their demand curve to the right. As long as the expected gain in revenue exceeds the marginal cost, firms will increase their advertising outlays. Thus firms in monopolistic competition are likely to have higher selling costs than firms in perfect competition and Figure 9–13 understates the actual price differential associated with product differentiation.

To the extent that advertising is actually informative, the social benefit may

match the cost. But much of the advertising undertaken by American industries is misleading, designed to attract consumers from other products that are equally, if not more, desirable. The Hotelling model suggests that there is actually very little product differentiation within an industry, despite manufacturers' claims to the contrary. Some industries, such as the tobacco industry, have escalated to extremely high levels of advertising outlay, which merely serve to maintain market shares for the individual participants.

Product Differentiation in Oligopoly

The model of market performance when products are differentiated can also be applied to oligopolistic behavior, that is, when firms are large in relation to the market. In this case, firms will attempt to reduce the elasticity of their individual demand curves or shift them rightward by real or perceived product differentiation. From another perspective, product differentiation serves to isolate the firm from its competitors and hence reduces the risk associated with price changes.

In many cases, price-fixing agreements, price leadership, and market-sharing arrangements have resulted in increased product competition with a concomitant growth in advertising for the industry as a whole. This in turn raises selling costs, which will eventually be passed on to consumers.

9–6 THE IMPACT OF MARKET STRUCTURE ON FACTOR PRICING AND EMPLOYMENT

Imperfect competition affects factor prices and employment in several ways. First, imperfect competition in product markets affects the demand for factors of production. In addition, firms that are large in relation to product market demand are often large in relation to the total demand for factors of production which are specialized with respect to a single industry and for which the supply is inelastic. For instance, the output decisions of General Motors have a significant impact on the market for automobiles, and in addition, its employment decisions have a significant impact on the market for automobile machinists. Consequently, General Motors cannot obtain all the labor it seeks at a fixed price, since a substantial increase in demand will drive the price up. A market with one buyer and many small sellers is called a **monopsony.** A market with a few large buyers and many small sellers is an **oligopsony.**

In the case of labor, unionization and collective bargaining has often developed as a reaction to oligopsony or monopsony. A market with a single buyer and a single seller is called a **bilateral monopoly.** In this section we examine models of factor pricing and employment in these alternative market structures.

Demand for Factors of Production in Imperfectly Competitive Product Markets

A firm maximizing profits will acquire resources as long as the addition to revenue is greater than the addition to cost. For a firm in imperfect competition, the addition to revenue associated with hiring one additional unit of a factor is the addition to output (the marginal product, MP_f) multiplied by the change in revenue per unit of additional output (the marginal revenue, MR), which is the **marginal revenue product,** MRP_f, that is,

$$MRP_f = MR \times MP_f$$

If the firm is small in relation to total *factor* demand, factor prices will be invariant to the quantity it purchases, so that the cost of hiring an additional unit of the factor is the factor price, p_f. Assuming that MRP_f decreases with employment of the factor, as shown in Figure 9-14, the maximum profit position will occur at N_m, where

$$MRP_f = p_f$$

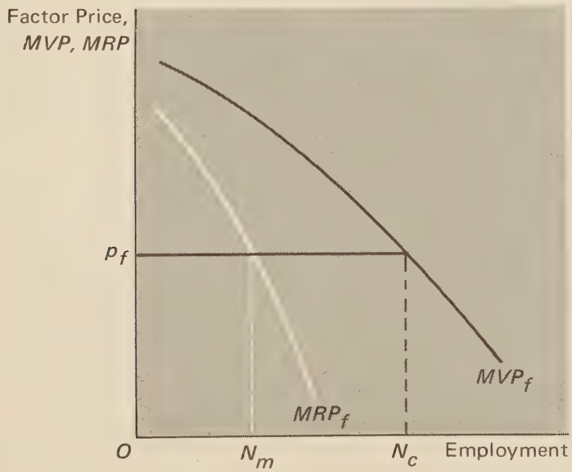


FIGURE 9-14
Demand for a Factor in an Imperfectly Competitive Product Market

Since marginal revenue is always less than price, marginal revenue product will always be less than marginal value product, that is,

$$p \times MP_f = MVP_f > MRP_f = MR \times MP_f$$

Consequently, for a given wage, employment in perfect competition, N_c, will be greater than in imperfect competition, N_m, as shown in Figure 9-14. Also, since

$$MRP_f = p_f < MVP_f$$

resources are paid less than their marginal value product.

Pareto efficiency requires that resources be paid in proportion to their marginal products. In addition, our analysis of income distribution using the concept of substitution elasticity assumed the marginal productivity principle. Does the existence of imperfect competition in product markets invalidate the marginal productivity principle of distribution?

For the individual firm, imperfect competition per se does not produce inefficient resource utilization for a given level of output. That is, firms will employ the most efficient factor combinations and produce at minimum cost for each level of output, although the output level itself will be below the Pareto efficient point where price equals marginal cost. To see why this is so, recall that efficiency requires the marginal rate of factor substitution to be equal to the ratio of the factor prices when the suppliers of those factors view the prices as fixed, that is,

$$MRS_{KL} = \frac{MP_L}{MP_K} = \frac{p_L}{p_K}$$

Since for firms in imperfect competition using both K and L,

$$MR \times MP_K = p_K \qquad \text{and} \qquad MR \times MP_L = p_L$$

then

$$\frac{\cancel{MR} \times MP_L}{\cancel{MR} \times MP_K} = \frac{p_L}{p_K} = MRS_{KL}$$

and an efficient combination of resources will be used by the firm in imperfect competition at each level of output.

Monopolistic Exploitation

Joan Robinson and others have argued that a factor that is paid less than the value of its marginal product is being exploited.[48] Clearly, this sort of "exploitation" will occur under imperfect competition. However, the difficulty could not be overcome by merely increasing the wage. In Chamberlin's model of monopolistic competition, for example, all factors are being exploited in the Robinsonian sense of being paid less than their marginal value products and yet all output is exhausted in factor payments and there are no excess profits. The problem is not only that factors are being paid less than their social value (to the extent that product price reflects social valuation) but that output levels in imperfect com-

[48] Joan Robinson, *The Economics of Imperfect Competition*, 2d ed. (London: Macmillan and Co., Ltd., 1969), chapter 25.

petition are below the social optimum (unless there are offsetting external diseconomies).

Impact on Factor Shares

The marginal productivity principle is often assumed to apply in analyzing the impact of changing technology or relative factor prices on income distribution. In fact, the principle is assumed to hold in empirical studies of the Unites States economy, despite the recognition that many markets are imperfectly competitive.[49]

Consider, for instance, the theoretical rationale for wage-price guidelines as an anti-inflation measure. If we assume the marginal productivity principle of wage determination, then product prices are established according to the formula

$$p = \frac{w}{MP_L}$$

where w is the wage and MP_L the marginal product of labor. As noted in chapter 7, a wage-price policy that permits wages to rise only as fast as productivity should have the effect of stabilizing prices. Furthermore, there would be no attempt by firms to boost their profits by raising prices, since the above formula represents the maximum profit equilibrium position and this position is unaffected by proportionate changes in wages and marginal labor productivity.

Suppose the industries that dominate economic activity are imperfectly competitive, setting

$$MR = \frac{w}{MP_L}$$

Recall that the relation between marginal revenue and price depends on the elasticity of demand, that is,

$$MR = p\left(1 - \frac{1}{E_D}\right)$$

so that

$$p\left(1 - \frac{1}{E_D}\right) = \frac{w}{MP_L}$$

As long as the elasticity of demand remains constant, then proportional changes in wages and productivity will have no impact on prices. Suppose, on the other hand, that businessmen perceive the elasticity of demand for their products as falling. This will increase the perceived degree of monopoly power, and firms will raise prices, even though marginal costs are unchanged. Changes

---

[49] See, for instance, Robert M. Solow, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, XXXIX (August 1957), pp. 312–320.

in the perceived degree of monopoly in the product market will alter the relation between marginal value product and marginal revenue product. For any individual firm, factors will be paid in proportion to their marginal value product, but among firms this will not be the case as long as the degree of monopoly varies among firms.

Demand for Factors of Production in Imperfectly Competitive Factor Markets

Consider the case in which a firm is not only a product market monopolist but also a factor market monopsonist.[50] In this situation, the firm's employment decisions have a significant impact on the input price if the factor is inelastically supplied. Figure 9–15 shows how the factor price and employment are established in a

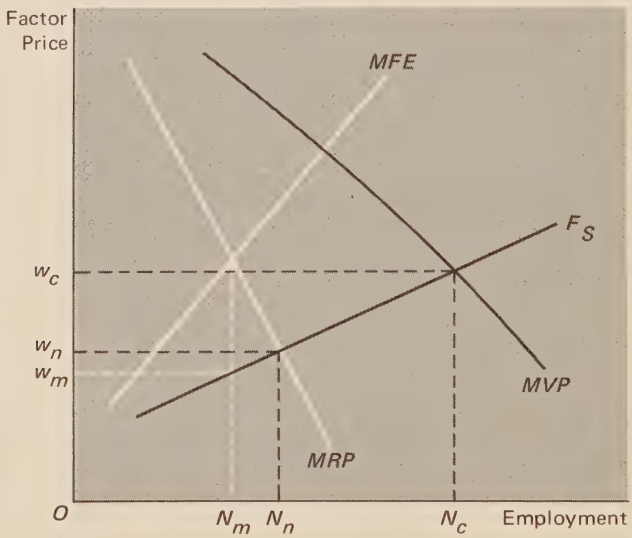


FIGURE 9–15
A Monopsonistic Factor Market

monopsonistic factor market. F_S is the market supply curve of the factor. *MFE* is the marginal factor expense or the change in total factor outlays associated with hiring more of a factor. If an increase in employment drives up the market price of the factor, then all units employed must receive the new, higher price and the marginal factor expense is greater than the factor price. Algebraically,

[50] The analysis of an oligopolist-oligopsonist is analogous, but account must be taken of rivals' possible reactions. For a discussion of some models of oligopsonistic market behavior see James M. Henderson and Richard E. Quandt, *Microeconomic Theory,* 2d ed. (New York: McGraw-Hill, 1971), pp. 242–243.

$$MFE = p_f + N\left(\frac{\Delta p_f}{\Delta N}\right)$$

that is, the price of an additional unit of the factor plus the change in factor price attributable to the acquisition of the new unit multiplied by all units employed is the marginal factor expense.

Profit maximization occurs where

$$MRP = MFE$$

at a level of employment of N_m and a factor price w_m. Notice that the wage, w_m, is determined from the factor supply curve, F_S, at N_m, since this is the wage required to induce N_m workers into the labor force. Monopsonistic equilibrium is below the competitive equilibrium (w_c, N_c) with respect to both wages and employment as well as below the equilibrium for the nonmonopsonistic monopolist with respect to both variables (w_n, N_n).

Monopsonistic exploitation is different from monopolistic exploitation, since the wage is below the marginal revenue product as well as below the marginal value product. Thus firms could conceivably be induced to increase wages if their monopsonistic power were reduced or eliminated.

Bilateral Monopoly

Recognition that monopsonistic exploitation can be reduced or eliminated by weakening a monopsonist's market power has served as an incentive to unionization of labor in such markets. In chapter 7 it was noted that a union can raise wages in a competitive labor market only at the expense of a reduction in employment unless it can effect an increase in product market demand. However, in a monopsonistic market a union can theoretically effect a wage increase without reducing, and possibly by increasing, the demand for labor. Consequently, unionization is likely to be most fruitful in monopsonistic or oligopsonistic labor markets.

Although there are many possible solutions to a bilateral monopoly bargain, depending on the objectives of the participants, their ability to collude, and other factors, we will limit our analysis to a noncollusive bargaining problem between labor and management. In Figure 9–16 the labor supply curve is L_S and the marginal factor expense curve is *MLE*. The monopsonist would hire N_m workers at a wage of w_m in the absence of countervailing union power. If the union could completely eliminate the monopsonist's factor market power, it could at the existing level of employment N_m, obtain an increase in wages to w_s. Alternatively, if employment were to increase from N_m to N_b, the union could obtain wage increases along MRP_L from just below w_s to w_b.

Consider two alternative strategies the union might use to achieve these results. Suppose it wishes to obtain the highest wage possible for those already employed. It could establish a completely inelastic supply curve at N_m by measures that would restrict labor supply to that number. This is shown as L_{Sr} in Figure

9–16. The equilibrium wage will rise to w_s with no reduction in the level of employment.[51]

Suppose, on the other hand, the union wishes to increase the level of employment as well as wages. In an extreme case, it could demand a wage of w_b for all members, resulting in a supply curve of w_bEL_S in Figure 9–16. N_b workers would be employed at the wage of w_b. By setting the wage between w_s and w_b, the union could effect increases in employment from N_m to N_b along the marginal revenue product curve.

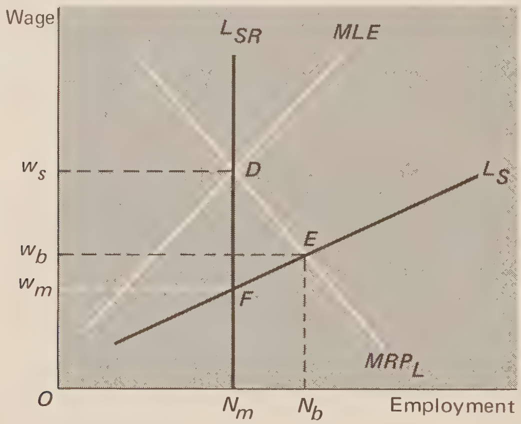


FIGURE 9–16
Wages and Employment Under Bilateral Monopoly

In the cases we have just described, the union is able to extract all monopsonistic power from the firm. In fact, of course, the firm recognizes that the supply price of labor is below the wage set by the union, and if the firm holds out long enough, it can induce workers to work at a lower wage regardless of union demands. Thus the actual bargain reached may result in wage payments below the marginal revenue product of labor, with an equilibrium reached somewhere in the area *DEF* in Figure 9–16. The outcome in any particular situation will depend upon the relative bargaining strengths of the bilateral monopolists.[52] The position of either side is weakened, of course, if there are other firms or unions in the bargain—the case of bilateral oligopoly. The position of the firm may also be weakened if it is a product market oligopolist. For instance, this may be the case for a firm with a monopsony in a regional labor market facing product competition from firms in other regions.

[51] Notice that the workers employed now earn a rent (in excess of the supply price of the last worker) equal to the amount of "exploitation" previously received by the firm.

[52] One way to gauge relative bargaining strength in a union-management negotiation is to evaluate the rate of the decline of goods inventories relative to the rate of depletion of strike funds.

QUESTIONS FOR STUDY AND REVIEW

1. "The models of perfect competition and monopoly are ideal types which are not intended to describe the actual behavior of firms (although some market structures may approximate these models)." Assuming this statement is true, what is the function of these models in economic theory and why do we bother to study them?
2. What is the theoretical justification for the "competitive ideal"? How is the validity of the argument affected by the phenomenon of product differentiation? How is it affected by the existence of economies of scale? What other characteristics of production might affect the argument?
3. A producer of electric power can separate its consumers into two groups, domestic and commercial users, and can charge different unit prices to the two groups. Assuming the firm is attempting to maximize profits, would there by any reason for it to charge different prices to the two groups? If so, which group would be charged a lower price, and why? What are the implications of your result for allocative efficiency?
4. Can a price ceiling imposed on a monopolist induce him to produce and sell more or less than he would have produced and sold without such limitation? Explain your answer in either case.
5. Antitrust policy in the United States has been based on the notion that market performance is related to market structure. In most cases, it is presumed than an industry composed of more than one or two firms performs better than if the industry were monopolized. State the reasoning on which such a policy is based and evaluate it critically.
6. "A firm maximizing profits will never produce in the inelastic range of its demand curve." Discuss.
7. Who will prefer a higher selling price for a book: the publisher or the author, who receives a royalty that is a fixed per cent of total sales?
8. Discuss and critically evaluate the welfare implications of Chamberlin's model of monopolistic competition.
9. "If workers are being paid a wage below their marginal value product, then unionization and collective bargaining could succeed in raising wages without a cutback in employment." Discuss.

ADDITIONAL READING

Arant, Willard. "Competition of the Few Among the Many," *Quarterly Journal of Economics, LXX* (August 1956), pp. 327–345.

Bain, Joe S. *International Differences in Industrial Structure.* New Haven: Yale University Press, 1966.

Baumol, William J. *Business Behavior, Value and Growth,* rev. ed. New York: Harcourt, Brace and World, 1967.

Chamberlin, Edward H. *The Theory of Monopolistic Competition,* 8th ed. Cambridge, Mass.: Harvard University Press, 1962.

Cyert, R. M., and J. G. March. *A Behavioral Theory of the Firm*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Hall, Robert and Charles Hitch. "Price Theory and Business Behavior," *Oxford Economic Papers,* No. 2 (May 1939), pp. 12–45.

Hotelling, Harold. "Stability in Competition," *Economic Journal, XXXIX* (March 1929), pp. 41–57.

Kaldor, Nicholas. "Market Imperfection and Excess Capacity," *Economica,* (n.s.) *II* (1935), pp. 33–50.

Machlup, Fritz. "Theories of the Firm: Marginalist, Behavioral, Managerial," *American Economic Review, LVII* (March 1967), pp. 1–33.

Mansfield, Edwin (ed.), *Monopoly Power and Economic Performance,* rev. ed. New York: W. W. Norton, 1968.

Modigliani, Franco. "New Developments on the Oligopoly Front," *Journal of Political Economy, LXVI* (June 1958), pp. 215–232.

Robinson, Joan. *The Economics of Imperfect Competition,* 2d ed. London: Macmillan and Co., Ltd., 1969.

Stigler, George. "The Kinky Oligopoly Demand Curve and Rigid Prices," *Journal of Political Economy, LV* (October 1947), pp. 432–449.

10 Resource Allocation over Time

Up to this point we have studied resource allocation without considering the passage of time. Factors of production are applied to processes that produce goods and services for instantaneous consumption. The household allocates its budget among goods and services, which are immediately consumed.

The question of when goods and services are consumed, and how long it should take to produce them, is just as important for the theory of resource allocation as what is consumed and how it is produced. Should a public program to help the poor take the form of direct welfare payments or aid to education? Although this question has elements of "what," it also implies a concern with when benefits are to be received. The benefits to education pay off over a longer time period than direct welfare payments.

A household may be faced with a choice between a $100 washing machine with a life expectancy of two years and a $250 model with a life expectancy of six years. Similarly, a firm may have to choose among different types of capital equipment that have different patterns of efficiency over time. One machine may produce uniform output for five years and then be worn out. Another may be more difficult to operate, so that in the first year, as the labor force is learning to use it, output may be low, rising in subsequent years. This machine lasts for ten years before wearing out. How does the firm choose between these options?

10–1 ALLOCATIVE EFFICIENCY OVER TIME

The criterion for the achievement of allocative efficiency over time is analogous to the criterion for static efficiency. The marginal rate of substitution of future for present consumption must be equal to the marginal rate of transformation of future for present production. That is,

$$MRS_{fp} = MRT_{fp}$$

In other words, current resources should be used for producing goods and services for future consumption until, at the margin, the amount that consumers are willing to forgo now for future consumption is equal to the opportunity cost of those future goods and services in terms of present consumption. This condition must hold for all goods and services that are produced by time-consuming techniques.

In order to evaluate whether or not this efficiency criterion is actually being achieved, it is important to understand the consequences of incorporating time into the production function. In chapter 6 we treated labor (a primary factor of production) and capital (a produced factor of production) as if there were no analytical differences between them. But, in fact, any process that uses capital is time-consuming and any theory of capitalist production must take time explicitly into account.

10–2 CAPITAL AS A FACTOR OF PRODUCTION

Factors of production other than labor and natural resources are capital, since they must first be produced by other resources. Even labor and natural resources contain elements of capital. Human capital, in the form of education, on-the-job training, and other acquired skills, is "produced" by the application of labor (and capital) at a previous time. Natural resources must be extracted from the earth, using labor and capital. Neoclassical capital theory viewed unskilled labor as the only primary (nonproduced) factor of production.

When we think of capital in a production process, we normally think of machines, blast furnaces, and buildings. However, another way to look at capital is as the accumulation of unskilled labor, applied at different points in time. Instead of viewing a blast furnace as a hunk of iron and steel, think of the labor that originally mined the ore and the coal for smelting the steel and the labor that fabricated the steel. Then the capital stock becomes a two-dimensional entity, measured in terms of unskilled labor input and the time at which the labor entered the production process. For instance,

$$K_t = 1l_t + 3l_{t-1} + 5l_{t-2} + 2l_{t-3} + 1l_{t-4}$$

where l_{t-i} refers to a unit of labor applied to the production of capital i years ago and K_t is the capital stock.

Clearly, we can increase the *amount* of capital by increasing the amount of labor applied at each period, for instance,

$$2K_t = 2l_t + 6l_{t-1} + 10l_{t-2} + 4l_{t-3} + 2l_{t-4}$$

This type of capital accumulation is called **capital widening.** Alternatively, we could increase the length of the production period, for instance,

$$2K_t = 1l_t + 4l_{t-1} + 7l_{t-2} + 4l_{t-3} + 3l_{t-4} + 2l_{t-5} + 1l_{t-6}$$

Increasing capital by increasing the length of the production period is called **capital deepening.**

A production process is said to be more **roundabout** if more labor is applied at an earlier period of time. More roundabout processes should require less total labor because they are more time-consuming. The first process requires four years from the original application of labor until the final output is obtained. The second process requires six years and ties up more labor in the earlier years than the first. Suppose that the capital produced by the first process is combined with primary labor to produce an output that is sold for $1,000. The capital produced by the second process is not available for two more years. During that time the proceeds of the sale can be invested in the bond market at interest, so that at the end of the two years they are worth more than $1,000. For the second to be feasible, it must be less costly in terms of primary resources than the first process. Alternatively, if a more roundabout process uses the same total amount of labor, it must be more productive than a less roundabout process. This is true even in the absence of an interest rate, provided consumers prefer immediate to future consumption.[1] Use of the less roundabout process will represent a gain in consumer satisfaction.

Roundabout processes are not necessarily more efficient than more direct techniques. However, a firm will never use a more roundabout technique unless it is either more productive or uses less total labor than a less roundabout process.

Capital and the Labor Theory of Value

The classical labor theory of value, developed by Ricardo, was a theory of relative cost. Later Marx transformed the theory into a normative model that related value to income distribution.[2]

In its naïve form the labor theory of value stated that commodities are exchanged in proportion to the labor required to produce them. This would be consistent with the observation that in a competitive market system commodity price ratios are equal to their marginal rates of transformation, if labor were the only cost of production. However, commodities produced with more capital-intensive techniques are clearly more costly for a given labor input than labor-intensive commodities.

Marx pointed out that capital is nothing but stored-up labor applied at a previous period. Consequently, capital-intensive products have a greater value because they contain more labor. Marx viewed the ratio of "dead" to "live" labor in a production process as a measure of its capital intensiveness.

Böhm-Bawerk, an Austrian economist, pointed out that the total amount of "dead" labor is not an accurate measure of the amount of capital in a production

[1] However, if there is a preference for immediate consumption, there will always be an interest rate in a market economy.

[2] A good summary of Marx's economic theory is Joan Robinson, *An Essay On Marxian Economics*, 2d ed. (New York: St. Martin's Press, 1966).

process, since the *time* at which the labor entered the process is also important.[3] Since more roundabout processes are more productive, the earlier labor is applied the more valuable it is. If the two processes require the same total amount of labor, the more roundabout one contains more capital.

The ensuing controversy between the Austrians and the Marxists was focused on a distributive issue. Marx had argued that since all value is derived from labor, all of the product rightfully belongs to labor. If capital produced by a more roundabout process is more productive than capital produced more directly but with the same total amount of labor, then to whom does the excess product belong? The Austrians argued that this is a payment to the capitalist for forwarding wages to workers whose labor was applied in the early stages. The more roundabout the process, the longer would the capitalist have money tied up in the "wages fund." This position was unacceptable to the Marxists, who felt that all value comes from labor.

Marxists today concede that Marx was in error when he ignored the importance of time in the production function.[4] As we will see, the optimal length of the production period is an important question for resource allocation. In a socialist economy, where such decisions are centrally planned and there is no capital market, the recognition of the role of time and the need to allocate resources over time is of crucial importance.

The Average Period of Production

Böhm-Bawerk developed a measure of the stock of capital that depends on the amount of labor used to produce it and the length of the production process. He defined the **average period of production** as a time-weighted average of labor input, that is,

$$p = \frac{L_t + 2L_{t-1} + \cdots + nL_{t-n}}{L_t + L_{t-1} + \cdots + L_{t-n}}$$

where p is the average period of production and L_{t-i} is the total amount of labor going into the production process in period i.

If labor is applied uniformly over time,[5] then

$$p = \tfrac{1}{2}t + \tfrac{1}{2}$$

[3] E. Böhm-Bawerk, *The Positive Theory of Capital*, trans. George D. Huncke and Hans F. Sennholz (South Holland, Ill.: Libertarian Press, 1959).

[4] Other Marxists claim that Marx never ignored the importance of time. In any event, Böhm-Bawerk's basic point is recognized as accurate.

[5] Since

$$1 + 2 + 3 + \cdots + t = \tfrac{1}{2}t(t + 1)$$

then

$$p = \frac{\tfrac{1}{2}t(t + 1)}{t} = \tfrac{1}{2}t + \tfrac{1}{2}$$

If labor is applied uniformly over 10 years, the average period of production is $5\frac{1}{2}$. If it is applied over 20 years, the average period is $10\frac{1}{2}$.

The formula

$$p = \tfrac{1}{2}t + \tfrac{1}{2}$$

does not allow for different speeds of accumulation. The same average period of production for different absolute periods of production, t, can be obtained by varying the amount of labor applied each year.[6]

Böhm-Bawerk's average period of production is a measure of the capital intensiveness of production, analogous to the capital output ratio in the modern development literature. It reflects the amount of "waiting" associated with a stipulated level of investment.

The average period of production as defined here is independent of the rate of interest. Clearly, the opportunity cost of waiting should be reflected in the value of capital. Another measure of the average period of production would weight each labor input by the interest rate, compounded since the time the labor was applied to the process, that is,

$$p' = \frac{L_t + (1+i)L_{t-1} + (1+i)^2 L_{t-2} + \cdots + (1+i)^n L_{t-n}}{L_t + L_{t-1} + L_{t-2} + \cdots + L_{t-n}}$$

Either measure, however, provides an ordinal index of capital intensiveness for comparing different equilibrium situations characterized by differences in the amount and structure of capital.

10–3 CAPITAL PRODUCTIVITY AND THE INTERNAL RATE OF RETURN

Suppose we wish to determine whether to allocate additional resources from present to future consumption. We would need to know the opportunity cost of future goods in terms of present goods, or the marginal rate of transformation of future for present goods. Assume that the labor force is fixed and fully employed, so that the only way to increase the amount of future goods is to deepen the capital stock. This means we must increase the degree of roundaboutness, or the average period of production. Böhm-Bawerk assumed that if the labor force remained fixed when the average period of production increased, output would increase, but at a decreasing rate, as shown in Figure 10–1. This assumption was consistent with the principle of diminishing returns, which characterized the neoclassical theory of production. As shown in Figure 10–2, increases in the average period of production result in a constantly decreasing marginal product.

Since in Böhm-Bawerk's model of stationary states capital accumulation can only take place if the average period of production is increased, this implies

[6] Böhm-Bawerk applied his analysis to a stationary state with zero population growth. The labor force remained fixed, so that it was necessarily applied uniformly over time. As we will see, this oversimplified assumption produced misleading results.

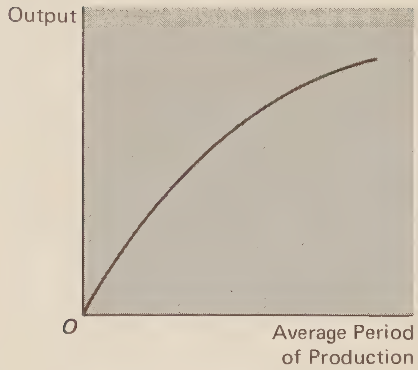


FIGURE 10–1
Output as a Function of the Average Period of Production

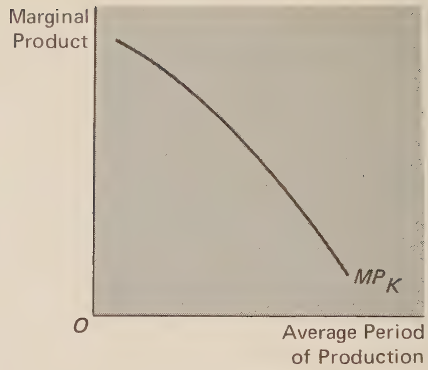


FIGURE 10–2
Marginal Product of Capital (Capital Deepening)

that the marginal productivity of capital decreases as the capital output ratio is increased, that is, as production becomes more capital-intensive. Furthermore, the marginal rate of transformation of future for present goods is 1 plus the change in output associated with increasing the period of production, that is,

$$MRT_{fp} = 1 + MP_K$$

where MP_K is the marginal product of capital. $-MRT_{fp}$ is the slope of the transformation curve between future and present goods. Since MRT_{fp} is a decreasing function of the period of production, as more resources are devoted to future consumption the slope of the transformation curve decreases in absolute value. Thus the transformation curve is convex to the origin, as shown in Figure 10–3.

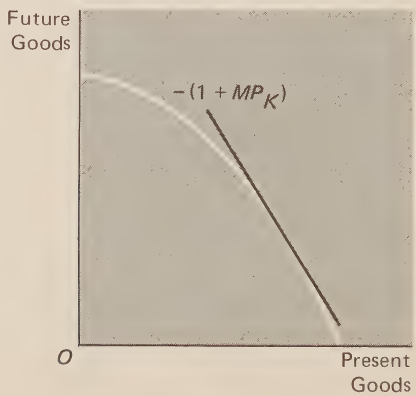


FIGURE 10–3
Transformation Curve Between Present and Future Goods

The Internal Rate of Return

Böhm-Bawerk's model of labor uniformly applied to produce a stationary level of output is not very useful for determining the opportunity cost of future goods in terms of present goods in a practical situation. For an individual firm, the capital stock can be increased by purchasing more capital goods at any one time as well as by lengthening the period of production. Thus capital accumulation has elements of widening as well as deepening.

Point Input, Point Output

Suppose a firm purchases a piece of capital equipment with the expectation of a single return after a certain number of years. The aging of wine is often cited as a point input, point output investment project.[7] The wine is purchased before it is fully aged, and at maturity it is sold.

Assume the initial cost of the investment is $1,000, with the expectation of a single return in one year of $1,100.[8] There are no other returns. Clearly, the rate of return on the investment, ρ, is 10 per cent, since

$$\rho = \frac{\$1{,}100 - \$1{,}000}{\$1{,}000} = \frac{100}{1{,}000} = 0.10$$

Point Input, Continuous Output

The typical investment project entails the purchase of capital equipment that yields a **stream of returns** over time. The returns stream can be uniform or nonuniform.

Suppose the firm purchases a machine for $1,000. It expects output to increase by 100 units each year as a result of using the new equipment. If output sells for $1 per unit, then the annual rate of return on the investment is 10 per cent, that is,

$$\rho = \frac{\$100}{\$1{,}000} = 0.10$$

But most investment projects do not generate a uniform stream of receipts over time. In general, efficiency increases in the beginning, as labor and management are learning to use the equipment, and then declines as the machine wears out. Decreasing returns can also occur if the firm is uncertain about prospects for future sales after a certain period. Estimation of future returns from investment

---

[7] Investment is another expression for capital accumulation, or change in the capital stock, that is,

$$I = \Delta K \qquad \text{or} \qquad I_t = K_t - K_{t-1}$$

[8] With reference to an individual firm, we can evaluate the productivity of capital in terms of dollars rather than labor units and units of physical output. This assumes that the monetary standard of value remains unchanged over the period.

requires projection into the unknown future. To guard against losses, firms will normally project returns at the level of minimum expectations. The farther into the future they project the greater the discount for risk and the lower the projected return.

Suppose a machine costs \$1,000 and yields a return of \$300 in the first year, \$400 in the second, \$400 in the third, \$100 in the fourth, and then wears out. The time pattern is shown in Figure 10–4. Can the "marginal productivity of capital" in terms of a compound rate of return be determined? Can this project be compared with one costing \$1,000 and earning a fixed annual rate of return forever?

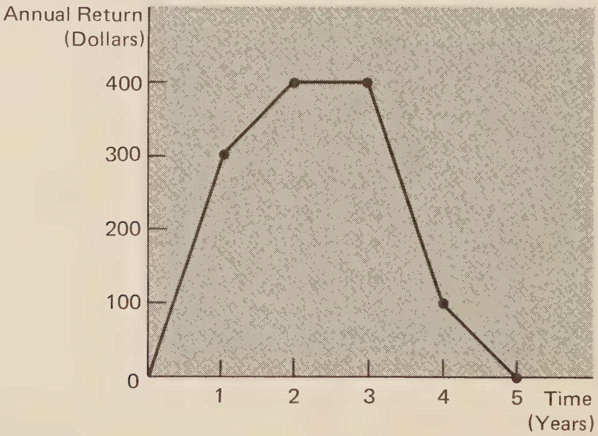


FIGURE 10–4
Nonuniform Returns: Point Input, Continuous Output

In the point input, point output case, the return in a single year could be computed by the formula

$$\rho = \frac{R - C}{C} = \frac{R}{C} - 1$$

or

$$C = \frac{R}{1 + \rho}$$

where R is the return and C is the initial outlay.

If i is the market rate of interest, then P, which is derived from the formula

$$P = \frac{R}{1 + i}$$

is called the **present value** of the asset.[9] Since financial assets earn compound

[9] The term *present value* is used because an asset yielding R dollars after one year is worth exactly as much as a financial asset costing P dollars this year.

interest, then to be equally attractive an investment project must also earn a compound rate of interest when the returns stream exceeds one year. Real investment must be at least as attractive as financial investment to induce firms to tie up their money in durable machines and equipment. One hundred dollars loaned at 10 per cent interest would be worth \$110 in one year, since

$$\$100 + (\$100 \times 0.10) = \$110$$

or

$$\$100(1 + 0.10) = \$110$$

In two years it will be worth not \$120 but \$121, since

$$\$100 + (\$100 \times 0.10) + (\$110 \times 0.10) = \$121$$

or

$$\$100(1 + 0.10)^2 = \$121$$

The present value of \$121 to be paid after two years is \$100, or

$$\$100 = \frac{\$121}{(1 + 0.10)^2}$$

Algebraically,

$$P = \frac{R_2}{(1 + 0.10)^2}$$

where R_2 is the return expected after two years.

Let the symbol R_j refer to a return expected after j years. Then the present value of an asset yielding a returns stream over n years is

$$P = \frac{R_i}{(1 + i)} + \frac{R_2}{(1 + i)^2} + \frac{R_3}{(1 + i)^3} + \cdots + \frac{R_n}{(1 + i)^n}$$

The compound rate of return on an investment project is that rate which equates the present value of the expected returns to the initial cost. Irving Fisher called this the **internal rate of return.**[10] If cost C is known, the compound rate of return, ρ, can be computed by the formula

$$C = \frac{R_1}{(1 + \rho)} + \frac{R_2}{(1 + \rho)^2} + \frac{R_3}{(1 + \rho)^3} + \cdots + \frac{R_n}{(1 + \rho)^n}$$

In our example, ρ can be computed from

$$\$1{,}000 = \frac{\$300}{(1 + \rho)} + \frac{\$400}{(1 + \rho)^2} + \frac{\$400}{(1 + \rho)^3} + \frac{\$100}{(1 + \rho)^4}$$

Solving by trial and error, we obtain

$$\rho \approx 0.085$$

[10] This is also Keynes's *marginal efficiency of capital* and Lerner's *marginal efficiency of investment.*

If the market rate of interest were 8.5 per cent, the firm would be indifferent between the real investment project and financial investment.

Continuous Input, Continuous Output

Suppose acquisition of the capital equipment entails cost outlays over a period of years. As long as the bulk of the returns occur later in time than the cost outlays (a reasonable assumption), the compound rate of return on the investment project is the rate that would equate the present value of the returns with the present value of the costs, that is,

$$\frac{R_1}{(1+\rho)}+\frac{R_2}{(1+\rho)^2}+\cdots+\frac{R_n}{(1+\rho)^n}=\frac{C_1}{(1+\rho)}+\frac{C_2}{(1+\rho)^2}+\cdots+\frac{C_n}{(1+\rho)^n}$$

If the bulk of the costs were incurred after the returns, ρ would measure the compound rate of cost over returns.

10–4 INVESTMENT CRITERIA AND THE OPTIMAL RATE OF INVESTMENT

The Austrian school addressed the question of the optimal level of investment in terms of the optimal period of production. This is because output could be increased only by varying the length of the production process.

Suppose a farmer is engaged in tree growing. As the trees age, the price he can obtain from the lumber yard increases. But every year he abstains from cutting his trees he forgoes the revenue from their sale, which he could invest at the market rate of interest. Consequently, he will only refrain from cutting his trees as long as the additional return associated with aging them one more year exceeds the market rate of interest. Assuming diminishing returns to aging, this implies that the optimal period of production occurs where the marginal rate of return to increasing roundaboutness, MP_K, equals the market rate of interest, i, that is,

$$MP_K = i$$

Changes in the Interest Rate

The Austrians argued that a decrease in the interest rate must increase roundaboutness and hence the capital intensiveness of production. They also concluded that the optimal capital-labor ratio is a decreasing function of the ratio of interest to real wages. That is, the higher the ratio of interest to real wages the lower will be the capital-labor ratio.

Both of these conclusions rest on the assumption that labor is uniformly applied over the period.[11] A decline in the interest rate reduces the opportunity cost

[11] These conclusions are also valid for the point input, point output case.

of waiting for future returns and hence encourages more roundabout processes. But it also reduces the initial costs of investment, making formerly unprofitable projects feasible.

Suppose project A has a large initial input of labor over two years, while B has a longer absolute period of production. B has a small initial input of labor and a large one at the end, which is, however, less than the initial input in A. Furthermore, the total input of labor in A is less than in B. At very high interest rates, compound interest on the labor at the start of method B outweighs the wage and interest costs of A, so A is preferred to B. As the interest rate falls relative to wages, the opportunity cost of waiting is reduced, and because the initial labor input is smaller in B, B is preferred to A. At low or zero interest rates, A is once again preferred, because it has a smaller total input of labor.[12]

Since the period of production as defined by Böhm-Bawerk is independent of the interest rate, it is clear that there will not be a one-to-one correspondence between the period of production and the interest rate when "reswitching" of techniques occurs as just described. There is a problem with defining capital accumulation strictly in terms of the period of production when inputs are not uniformly applied.

The Present Value Criterion

Suppose a firm is considering investing $1,000 in expanding its capital facilities. The firm also has the option of buying a bond that guarantees a return of 10 per cent. Thus the opportunity cost of the investment project is the market rate of interest, or 10 per cent. The present value of a bond earning the market rate of interest is its initial cost, in this case $1,000. The firm will undertake the investment project in preference to purchasing the bond as long as its present value exceeds the initial cost of $1,000. Present value is a way of expressing all future returns in terms of their current worth, taking into account the returns that could have been made by investing the money at the market rate of interest. Thus a firm will undertake an investment project as long as

$$PV = \sum_{j=1}^{n} \frac{R_j}{(1+i)^j} > C$$

where $\sum_{j=i}^{n}$ means that the expression is a sum of terms where $j = 1, 2, \ldots, n$.

Suppose we consider all potential investment projects and rank them in descending order of their net present values. Net present value is the difference between present value and cost, that is,

$$NPV = PV - C$$

As long as the firm can borrow indefinitely at the market rate of interest (or has

[12] This example is taken from G. C. Harcourt, "Some Cambridge Controversies in the Theory of Capital," *Journal of Economic Literature, VII* (June 1969), pp. 393–394.

sufficient internal resources) it will always maximize profits by maximizing net present value.[13] When projects are mutually exclusive (for instance, a parking lot versus a hotel on the same piece of land), the one with the highest net present value is chosen.

A schedule of investment projects and their net present values is shown in Figure 10–5. If the firm's objective is to maximize profits, the equilibrium level of investment is I^*, where net present value is zero.

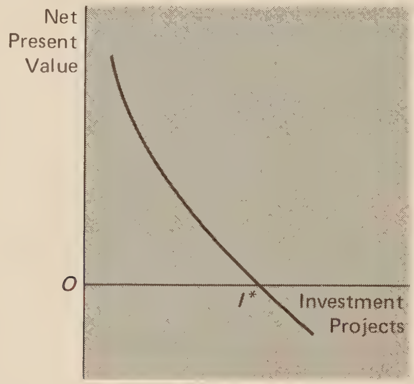


FIGURE 10–5
Profit-Maximizing Level of Investment Using Present Value Criterion

Effect of Interest Rate Changes

Changes in the rate of interest will affect present value and the optimal level of investment. A decline in the rate of interest will increase net present value for all projects, shifting the *NPV* schedule to the right and increasing the equilibrium level of investment.

In addition, a reduction in the interest rate is more favorable to projects with longer periods of production, since the opportunity cost of waiting is reduced. Projects with longer gestation periods experience greater increases in *NPV* than short-lived projects. Where projects are mutually exclusive, longer projects with lower *NPV* at high interest rates now become favored over projects with shorter gestation periods. Thus the lower the general level of interest rates, the more lengthy will be the type of investment projects chosen.

This observation is important for planning social investment. Suppose Congress has budgeted $100,000 for aid to a poverty area. Two projects are being considered: construction of new schools versus direct welfare payments to the

[13] Some planners have suggested maximizing the ratio between *NPV* and cost. The choice is similar to maximizing the *profit rate* versus maximizing *total profits*. High profit rates do not necessarily mean high total profits if total sales are small. A small project may have a higher profit rate than a larger one that has higher total profits.

poor. It is estimated that the benefits of the welfare program can be measured by the money value of the payments to welfare recipients. The benefits associated with the school construction program accrue over a period of twenty-five years. The *NPV* of the welfare program is zero. The *NPV* of school construction will vary with the rate of interest, being higher the lower is the rate of interest. At high interest rates the welfare program may be preferred, while at lower interest rates the school construction program may have higher *NPV*.

Since the choice of investment projects depends on the interest rate, social planners must be concerned with the question: What is the appropriate rate of interest for determining the present value of public investment projects? Many economists argue that existing market rates reflect the preferences of the current adult generation for present over future goods. The stronger the preference, the greater the rate of interest required to induce people to part with current resources. The market rate must understate social preference for future consumption, which includes the preferences of the young as well as future generations. Consequently, the social interest rate used for discounting public investment projects to their present value should be lower than the private market rate of interest. This will encourage investment in programs with payoff streams reaching far into the future—the space program, environmental protection, and so on—would never be undertaken on the basis of intertemporal preferences of the current adult generation.[14]

The Internal Rate of Return Criterion

Irving Fisher and later John Maynard Keynes were both interested in showing a more direct relation between the rate of interest and the level of investment.[15] Fisher was interested in the microeconomic theory of intertemporal resource allocation and a theory of interest, while Keynes was concerned with the macroeconomic problem of stabilizing investment through the use of monetary policy.

Both Fisher and Keynes took up where Böhm-Bawerk had left off. Assuming diminishing returns to investment, firms will accumulate capital until

$$MP_K = i$$

where i is the market rate of interest. The market rate of interest represents the opportunity cost of real investment in terms of financial investment.

If cost and return streams are not uniform, the internal rate of return, ρ, represents the compound rate of return on the investment project. Firms will undertake investment projects as long as ρ is greater than the interest rate. The

---

[14] For a discussion of using an interest rate other than the market rate of interest to determine the optimal period of social investment see William J. Baumol, "On the Social Rate of Discount," *American Economic Review, LVIII* (September 1968), pp. 788–802.

[15] Irving Fisher, *The Theory of Interest* (London: Macmillan and Co., Ltd., 1930), and John Maynard Keynes, *The General Theory of Employment, Interest and Money* (New York: Harcourt Brace Jovanovich, 1965).

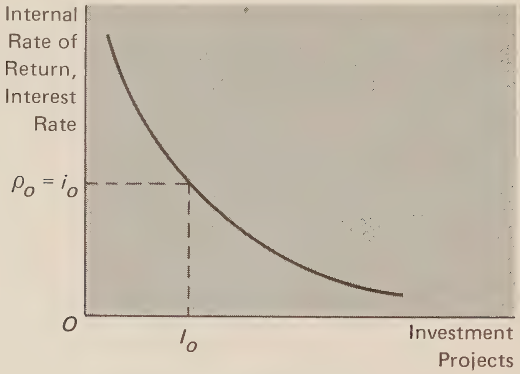


FIGURE 10–6
Profit-Maximizing Level of Investment Using Internal Rate of Return Criterion

internal rate of return criterion is equivalent to the net present value criterion, since all projects for which $\rho > i$ will have positive net present value, that is,

$$NPV > 0$$

This can be seen in the following way for the point input, continuous output case.[16] The internal rate of return, ρ, is the interest rate for which NPV is zero, that is,

$$\sum_{j=1}^{n} \frac{R_j}{(1+\rho)^j} = C$$

If i is greater than ρ, the present value of the returns would decline, making net present value negative, that is,

$$\sum_{j=1}^{n} \frac{R_j}{(1+i)^j} < C$$

for $\rho < i$.

If i is less than ρ, the present value of the returns would increase and net present value would be positive, that is,

$$\sum_{j=1}^{n} \frac{R_j}{(1+i)^j} > C$$

when $\rho > i$.

When ρ is equal to i,

$$\sum_{j=1}^{n} \frac{R_j}{(1+i)^j} = C$$

and net present value is zero.

[16] It is true for the general continuous input, continuous output case provided that the bulk of the costs are incurred before all returns are realized.

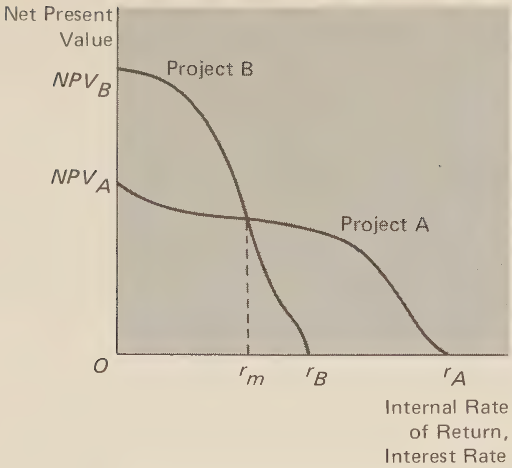


FIGURE 10–7
Investment Criteria for Mutually Exclusive Projects

Suppose all investment projects were ranked in order of their internal rates of return. As long as there were no mutually exclusive projects, investment would be undertaken until $\rho = i$. In Figure 10–6, when the interest rate is i_o, investment is I_o. A decline in the interest rate would increase the number of profitable investment projects.

Mutually Exclusive Projects

The internal rate of return criterion is not valid for ranking mutually exclusive projects that have different time profiles of cost and returns.

Consider two projects A and B that are mutually exclusive. A has a shorter returns stream than B and consequently its net present value, NVP_A, is less sensitive to changes in the market rate of interest than NVP_B. That is, changes in i cause greater changes in NVP_B than in NVP_A. The relation between the interest rate and NVP for the two projects is shown in Figure 10–7.

The internal rate of return is that interest rate for which NVP is zero. The internal rate of return on A, r_A, is greater than the internal rate of return on B, r_B. Thus, if A and B were mutually exclusive, A would be preferred to B by the internal rate of return criterion. But if the market rate of interest is below r_m (and both projects would be profitable under the internal rate of return criterion), the net present value of B is greater than NVP_A and B should be undertaken.[17]

[17] The fact that A is preferred to B at some rates of interest and B to A at others is analogous to the phenomenon of "switching techniques" in the production of capital goods discussed earlier.

The internal rate of return criterion is valid for determining if a project is profitable but is not a valid criterion for ranking mutually exclusive projects that are both profitable. For this purpose, the present value criterion must be used.[18]

Fisher suggested that mutually exclusive projects could be ranked by the internal rate of return criterion if one project were viewed as the opportunity cost of the other. Suppose project A is the opportunity cost of B. Fisher defines the **marginal rate of return over cost,** μ, as the hypothetical interest rate that would equalize the net present values of the two projects, that is,

$$\sum_{j=1}^{n} \frac{R_{jB}}{(1+\mu)^j} - C_B = \sum_{j=1}^{n} \frac{R_{jA}}{(1+\mu)^j} - C_A$$

If μ is greater than the interest rate, project B should be undertaken. If μ is less than the interest rate, A is more profitable.

In Figure 10–7 the marginal rate of return over cost is r_m. At interest rates below r_m, B is preferred. For this method to produce correct results, the short-lived project must always be viewed as the opportunity cost of the long-lived one. Otherwise, μ measures the rate of cost over return.

Investment, the Interest Rate, and Uncertainty

Both the Austrian and Fisherian approaches concluded that the level of investment is inversely related to the interest rate. But the Fisherian approach has more to say about the effects of the interest rate on the structure of investment projects. This is particularly important in a planning context in which funds for capital expenditures are fixed, so that all investment projects must be viewed as mutually exclusive. That is, there may be many investment projects that have internal rates of return or marginal productivities in excess of the interest rate but that cannot all be undertaken because of restricted funds. Many governments are constrained by debt ceilings imposed by legislatures, so that unlimited borrowing at the market rate of interest is not feasible. Firms also cannot usually borrow unlimited amounts at fixed interest rates.[19]

If the interest rate reflects society's preference for present over future consumption, then one would say that limiting investment below the level where the opportunity cost of future in terms of present consumption equals the interest rate produces an inefficient intertemporal allocation of resources. However, this analysis has failed to take into account the effect of uncertainty on investment decisions. In most cases, it is difficult to estimate the future time stream of

[18] For a further discussion see Armen A. Alchian, "The Rate of Interest, Fisher's Rate of Return over Cost, and Keynes' Internal Rate of Return," *American Economic Review, XLV* (December 1955), pp. 938–943.

[19] For a discussion of the relation between the level of investment and the cost of capital to individual firms see James S. Duesenberry, *Business Cycles and Economic Growth* (New York: McGraw-Hill, 1958), chapter 5.

benefits from a particular project, in view of rapidly changing tastes and technology. A hydroelectric power project may become obsolete if the use of nuclear power becomes widespread. Migration out of a poverty area may result in unused educational facilities. In addition, for public investment projects, the benefits are often difficult to evaluate in monetary units.[20] Consequently, the estimation of the social internal rate of return may be extremely arbitrary and subject to political manipulation.

Private firms can more easily quantify the returns derived from investment projects, since the goods they produce are sold to the public.[21] However, uncertainties related to changing tastes and technology may induce them to use a **payoff period** criterion for investment. According to this criterion, if the returns are not likely to exceed the costs after a stipulated number of years (depending on the unpredictability of changes in tastes and technology) the project should not be undertaken, regardless of the rate of interest. This approach obviously favors short-lived projects, and changes in the rate of interest may have little impact on investment decisions.

10–5 THE RATE OF TIME PREFERENCE

Our theory of the household assumed that consumers are concerned only with allocating their incomes among various goods and services. But the typical household must also consider the time pattern of these expenditures. Should a new car be purchased now or next year? Should we borrow money to send Jane to college and repay out of future income, or should we reduce our current level of consumption and finance her education from current resources?

Assume that relative prices of goods and services remain fixed over the period studied and that there are no changes in tastes. Thus we abstract from the problem of changes in the choice between commodities and concentrate on the time pattern of overall consumption.[22] Also assume that the household knows its future income with certainty and that it is willing to enter the bond market as either a borrower or a lender.

Figure 10–8 shows a set of consumer indifference curves for future consumption versus present consumption. Each curve is a locus of combinations of consumption next year (C_2) and consumption this year (C_1) that the household finds

---

[20] See Robert Dorfman (ed.), *Measuring Benefits of Government Investments* (Washington, D.C.: The Brookings Institution, 1965) for a discussion of some of the methods economists use to evaluate public investment programs.

[21] Private firms may misestimate social benefits and costs associated with their own investment projects if externalities are present. Externalities are not computed in the private internal rate of return.

[22] Obviously, the time pattern of consumption of some items such as automobiles and washing machines can be varied much more in response to changes in underlying conditions than consumption of others such as food.

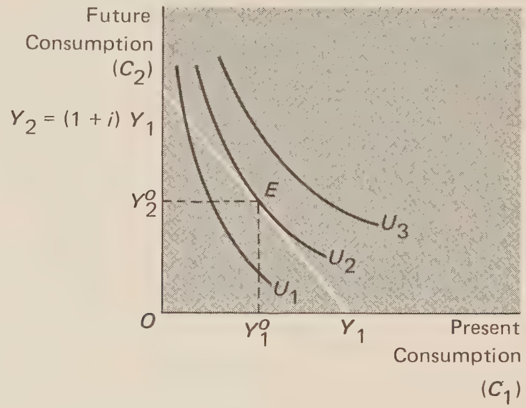


FIGURE 10–8
Household Indifference Curves for Future Consumption Versus Present Consumption

equally satisfactory.[23] The absolute value of the slope of each indifference curve is the marginal rate of substitution of future for present consumption, that is,

$$\frac{-\Delta C_2}{\Delta C_1} = MRS_{fp} = \frac{MU_{C1}}{MU_{C2}}$$

The MRS_{fp} is called the **rate of time preference.** Along any indifference curve, the more current consumption is given up for future consumption the higher is the rate of time preference, so the indifference curves are convex to the origin. Also, for the average household, we usually assume that if consumption is equally distributed between the two periods, the rate of time preference will exceed 1. That is, there is a preference for present consumption, and if people could borrow at a zero interest rate, they would consume a disproportionate share of their total lifetime income in early years. This implies that along a forty-five-degree line from the origin the slopes of the indifference curves would exceed 1 in absolute value.

It is also likely that the intertemporal preference function is not homogeneous. For any given distribution of consumption between two periods (along a straight line from the origin), time preference is likely to decline the higher the level of income. At very low levels of income, when subsistence needs are barely met, there is likely to be a greater underestimation of the future and stronger time preference than at higher levels. Thus the indifference curves become flatter farther from the origin.[24]

---

[23] The vertical axis could measure a perpetual (constant) future income stream as well. The two-period analysis is somewhat less complicated and can be generalized into the n-period case.

[24] This is not a necessary condition for an indifference map for an individual consumer, but it does seem reasonable from a theoretical point of view.

The budget line for the consumer is also shown in Figure 10–8. Y_1 represents resources available for consumption in the initial period. Suppose y_1 is expected income in period 1 and y_2 is expected income in period 2. Then Y_1 is equal to Y_1 plus the present value of y_2, that is,

$$Y_1 = y_1 + \frac{y_2}{1+i}$$

Y_2 stands for maximum total resources available in period 2: y_2 plus y_1 plus the interest earned by placing y_1 in the bond market for a year, that is,

$$Y_2 = y_2 + y_1(1+i)$$

Notice that

$$Y_2 = (1+i)Y_1$$

so that the absolute value of the slope of the budget line is

$$\frac{\Delta C_2}{\Delta C_1} = \frac{\Delta Y_2}{\Delta Y_1} = 1 + i$$

The household maximizes utility where the budget line intersects the highest indifference curve. This occurs at E, where the budget line is tangent to an indifference curve. Since they both have the same slope at the point of tangency

$$MRS_{fp} = 1 + i$$

A consumer maximizing utility will arrange his consumption over time, so that the marginal rate of substitution of future for present consumption is equal to 1 plus the market rate of interest.

Effect of an Interest Rate Change

Suppose there is an increase in the interest rate. This will reduce the present value of y_2 and increase the future value of y_1. The budget line shifts from AE to DB, as shown in Figure 10–9, with the point of intersection occuring at (y_2, y_1). At this point there is no lending or borrowing and there is no effect on the budget of a change in the interest rate.

Whether the consumer is better or worse off as a result of the increase in i will depend upon whether he was initially a borrower or a lender. If the consumer were initially at equilibrium to the left of (y_2, y_1), he is consuming less than his current income and is a lender. The rise in the interest rate increases his total budget and his new equilibrium would be on a higher indifference curve. If the initial equilibrium were to the right of (y_2, y_1) the household is consuming more than its current income and is a borrower. In this case, the rise in interest rates reduces the amount it can borrow on y_2 and consequently reduces total income. The new equilibrium would be on a lower indifference curve.

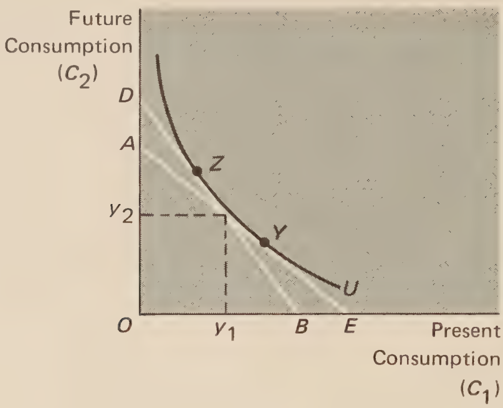


FIGURE 10–9
Change in the Market Rate of Interest

A rise in the interest rate always has a substitution effect that favors postponing present for future consumption. Consider the case shown in Figure 10–9. Initially, equilibrium is at *Y* and the household is a net borrower. The rise in the interest rate shifted the budget line to *DB* and the household became a net lender. The losses associated with the decline in the present value of y_2 are exactly offset by gains in the future value of y_1. Consequently, there is no income effect associated with the increase in interest rates. The substitution effect induces the household to reduce the level of current consumption.

It is generally assumed that a rise in interest rates will increase the rate of saving and reduce the level of current consumption for the economy as a whole. Theoretically this would occur if the income effect is ignored altogether. Since for every borrower there must be a lender, we might assume that the negative effects on borrowers in the aggregate would be exactly offset by gains to lenders, so that income effects will cancel each other out. However, the existence of a large public debt makes the private sector (households and businesses) net creditors. This means that a rise in interest rates will produce a net increase in income to the private sector (although not necessarily to an individual household), which may actually stimulate current consumption through the income effect. Although a rise in the interest rate will also stimulate saving (since both the substitution and income effects are favorable to saving), consumption may increase too.

Studies of aggregate consumption behavior show that changes in interest rates have little or no impact or aggregate consumption. This suggests that substitution and income effects may be working in opposite directions to cancel each other out. On the other hand, it is generally agreed that higher interest rates, other things being equal, are conducive to higher levels of aggregate saving.

10-6 DETERMINATION OF THE RATE OF INTEREST

The theory of intertemporal consumption can be combined with the theory of investment to show how the interest rate is determined in a market economy. There are a number of approaches to the theory of interest that are not necessarily contradictory but essentially designed for different purposes. Keynes, who was interested in the interest rate as a macroeconomic stabilizer, used a money market approach, which focuses on how changes in the money supply affect the interest rate. In the context of microeconomic theory, we are concerned with the interest rate as a "price" reflecting the scarcity of current resources. For our purposes, the interest rate is the regulator of intertemporal resource allocation just as product and factor prices regulate resource allocation within a period of time.

The Market for Loanable Funds

One way to look at the establishment of an equilibrium rate of interest is in the context of a market for loanable funds. The demand for loanable funds comes from firms who wish to invest in capital projects and from households who wish to consume more than their current income. In both cases, the demand for loanable funds will increase with a decline in the interest rate. Firms will find more investment projects profitable at lower interest rates. Also, lower interest rates will cause families to increase levels of current consumption through the substitution effect.

The supply of loanable funds comes from household savings and from surplus funds of firms that cannot find profitable investment projects at the prevailing rate of interest.[25] The supply of loanable funds increases with the rate of interest. As interest rates increase, firms may find financial investment preferable to real investment. In addition, rising interest rates stimulate saving through both the substitution and income effects.

The equilibrium rate of interest is determined in the market for loanable funds as shown in Figure 10–10. The government can manipulate the rate of interest by affecting the demand for, or supply of, loanable funds through open-market sales of government bonds or other monetary instruments.[26] Changes in consumers' time preferences and changes in the profitability of investment also affect the interest rate. An increase in time preference, for instance, will reduce the supply of loanable funds, shifting S to the left and causing an increase in the interest rate. An increase in government spending, by increasing aggregate demand, may increase the expected returns from investment. This will cause an

---

[25] In the American economy it is rare for firms to undertake financial investment with surplus funds. Studies suggest that firms generally view real investment opportunities as superior to financial investment. In the United States the household sector is a net lender while the business sector is a net borrower.

[26] For a discussion of the way government can affect the interest rate see *The Federal Reserve System* (Washington, D.C.: Board of Governors of the Federal Reserve System, 1965).

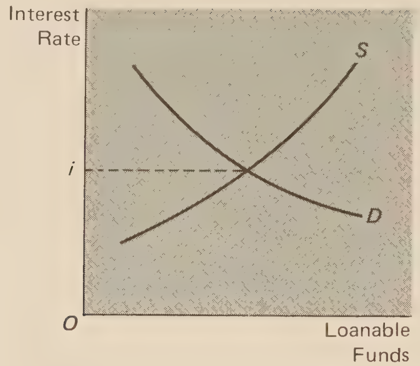


FIGURE 10–10
Determination of the Interest Rate in the Market for Loanable Funds

increase in the demand for loanable funds, shifting D to the right, and the interest rate will rise.

The Interest Rate and Economic Efficiency

Households maximizing utility will arrange their intertemporal consumption patterns so that

$$MRS_{fp} = 1 + i$$

Firms will undertake new investment projects until

$$1 + \rho = MRT_{fp} = 1 + i$$

Thus the establishment of an equilibrium rate of interest ensures that

$$MRS_{fp} = 1 + i = MRT_{fp}$$

which is the condition for intertemporal allocative efficiency. The implications of this result can also be seen graphically. Figure 10–11 shows a social transformation curve between future and present commodities. All potential investment projects are included and ranked in order of their net present value. The absolute value of the slope of the curve is the marginal rate of transformation of future for present commodities, $1 + \rho$, where ρ is the internal rate of return. The absolute value of the slope of the budget line AB is $1 + i$, where i is the rate of interest. Firms will invest $C_1^S E$ of current resources for projects that will pay off C_2^S next year.

W is an indifference curve from a preference function that represents the collective preferences of all households for future versus present consumption.[27]

[27] Income distribution is taken as given, so we have used the term collective preferences rather than social preferences.

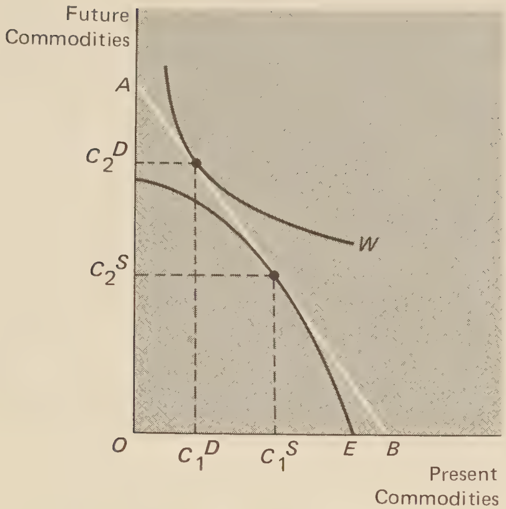


FIGURE 10–11
Disequilibrium Rate of Interest

Consumer equilibrium occurs where W is tangent to the budget line. Households will consume C_1^D of their current income and consequently will be willing to release C_1^DE of current resources for investment.

Clearly, the supply of loanable resources, C_1^DE, is greater than the demand, C_1^SE. Thus the interest rate implicit in the slope of AB is too high to clear the market. Lenders willing to accept lower rates will bid down the interest rate until the demand for loanable resources (funds) equals the supply.

Equilibrium in the market for loanable resources or funds occurs where the budget line is tangent to the intertemporal consumption indifference curve, W, and to the transformation curve simultaneously. Here W and the transformation curve are tangent to each other, as seen in Figure 10–12. Equilibrium occurs at the point of intersection between the transformation curve and the highest collective indifference curve on the preference map. Although the point does not necessarily maximize *social* welfare (since income distribution is taken as fixed), it is a Pareto efficient equilibrium. Thus, under the conditions described, a market system produces an efficient intertemporal allocation of resources, with the market rate of interest being the price that regulates the allocative process.

Market Failures

There are several cases in which a market system fails to produce optimal intertemporal resource allocation. First, as noted, the supply of loanable funds depends primarily on the subjective time preference of the present adult generation, which tends to undervalue the future. Furthermore, when income distribution is taken into account, investment decisions may be largely determined by the pref-

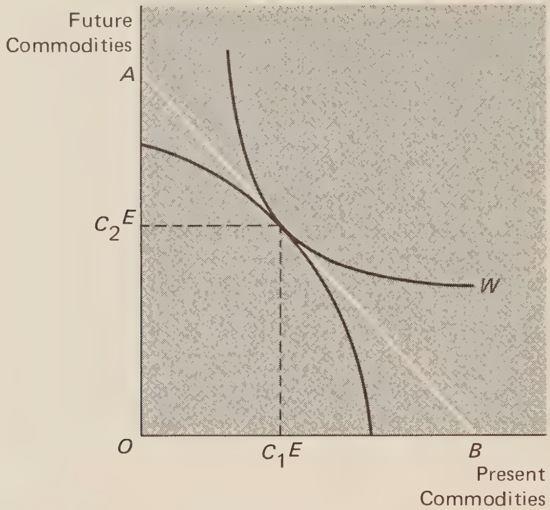


FIGURE 10–12
Equilibrium Rate of Interest

erences of specific individuals. This is particularly true when the benefits affect only a small group of people. The people who control the wealth can also determine how investment funds can be used—certain types of education are more highly subsidized than others, housing developments follow patterns reflecting middle-class values, and so on.[28] Because the benefits associated with many investment projects are intangible, and because of the elements of risk and uncertainty, decisions concerning the optimal investment project(s) depend more on the subjective evaluation of individuals than the theory suggests. Consequently, the initial distribution of wealth is an important determinant of the type of investment that actually takes place.

External economies and diseconomies associated with investment will not be taken into account by private firms. Automobile manufacturers do not allow for the impact on the environment in calculating the internal rate of return on their investment. Consequently, there may be excessive investment in that industry from the viewpoint of society as a whole. On the other hand, individuals may underestimate the gains from investing in education. Most studies of the internal rate of return to education measure the return solely in terms of the gain in lifetime income attributable to the added education.[29] There may, however, be social benefits not accounted for in market earnings. If so, individuals who pay the cost of their own education will underinvest in this form of human capital. In recogni-

[28] Wealthy people may have weaker time preference than the rest of the population, however, which may help to counterbalance the underestimation of the future by the present generation.

[29] For an example of such a study see Giora Hanoch, "An Economic Analysis of Earnings and Schooling," *Journal of Human Resources, II* (Summer 1967), pp. 310–329. The same issue contains a symposium on returns to investment in education.

tion of this, most governments have undertaken programs to subsidize education.

In all actual capitalist economies imperfections exist in the market for loanable funds. Most individuals are not able to borrow and lend at the same rates. Furthermore, certain individuals and firms are able to borrow at lower rates than others. The result is that the subjective rate of time preference of households is actually lower than the internal rate of return to investment. This implies that a market system with imperfect capital markets will have a lower rate of capital accumulation than is consistent with intertemporal allocative efficiency. Furthermore, when firms borrow at different rates of interest they invest at different internal rates of return. Thus the opportunity cost of investment is not uniform throughout the economy, which also represents a departure from the efficiency conditions.[30]

10–7 PLANNING FOR INTERTEMPORAL EFFICIENCY

Because of the pervasiveness of Marxist ideology in today's centrally planned economies, the question of intertemporal resource allocation has been the subject of considerable confusion. Marxist theory implies that since all value is attributable to labor, the interest rate is simply a means by which capitalists derive a share of the product that rightfully belongs to labor. Thus, in a Marxist state (after the revolution of the working people), there would be no interest.

The absence of interest implies, however, that capital is a free good. The interest rate is the price that indicates the opportunity cost of capital accumulation in terms of present resources. If the interest rate is zero, the opportunity cost must also be zero, which means that capital can be obtained with no sacrifice of current consumption. This is, of course, impossible if all resources are fully employed.

If there were no interest rate, there would be no incentive to postpone present consumption for the future, assuming a positive rate of time preference. Furthermore, there would be no incentive to replace worn-out equipment, so that the existing capital stock would eventually become depleted. Socialist planners could overcome these difficulties by forcing saving through taxation and using tax revenues for investment. However, in the absence of an interest rate there would be no way to evaluate alternative investment projects. Such an evaluation requires the computation of present value, which requires knowledge of an interest rate. How would a planner choose between a hydroelectric project that will begin to generate large quantities of power immediately versus a nuclear power plant that will take longer to construct but will generate even greater capacity in the future? The answer to this question requires knowledge of the social rate of time preference.

---

[30] For a further discussion see J. Hirshleifer, "On the Theory of the Optimal Investment Decision," *Journal of Political Economy, LXVI* (August 1958), pp. 329–352.

Oskar Lange, in his model of decentralized socialism, advocated interest charges on capital. Many of the Eastern European countries outside the Soviet Union have adopted the practice. However, by the late 1960s, the USSR had still refused formally to adopt the use of interest rates in its capital budgets, despite the open objections of some of its more outspoken academic economists.

A centrally planned economy would have the advantage of setting the interest rate at a level that reflects the social rate of time preference. Although any estimate of the social rate of time preference by the Central Planning Board would necessarily be arbitrary, theory suggests that the social rate of time preference should be lower than the market rate currently in effect in most capitalist economies. Setting the borrowing rate anywhere from 2 per cent to 4 per cent would probably be more consistent with social preferences than the somewhat higher rates that exist in market economies.[31] Since the private rate of time preference is likely to be higher than the social rate, sufficient funds are not likely to be forthcoming from private saving to meet the demand for loanable funds at the established rate of interest. Consequently, the balance would have to be derived from tax revenues.

An additional advantage to central planning is that a uniform rate of interest would be charged to all borrowers. This will ensure that the opportunity cost of capital accumulation is equalized throughout the economy.

QUESTIONS FOR STUDY AND REVIEW

1. Consider two investment projects yielding the income streams shown below:

Time Realized	R_1	R_2
1	$ 0	$300
2	$500	$300
3	$500	$300
4	$200	$300

   a) If the initial cost of each project is $1,000, what is the internal rate of return?
   b) If the market rate of interest is 5 per cent and the projects are not mutually exclusive, which should be undertaken?
   c) If the projects are mutually exclusive, which should be undertaken?
   d) How would your answer to (c) be affected if the interest rate were reduced to 2 per cent?

2. An economics student is considering two alternative job offers. Both involve research for professors, and the student finds them equally attractive. One pays $1,000 per year

[31] Nominal interest rates are higher than real rates of interest if inflation causes the value of money to depreciate over time. Thus, if prices rise by 5 per cent a year, a nominal interest rate of 8 per cent represents a real return of only 3 per cent. This must be kept in mind when assessing the level of real interest rates in economies experiencing price inflation.

for two years. The other pays $1,200 in the first year and $700 in the second year. For what rate of interest would the student be indifferent to the two jobs?

3. The term *value* in economics has meant different things to different people. Suppose we use *value* to refer to the opportunity cost of goods and services in terms of each other. Is the Marxian labor theory a good explanation of this kind of value? Explain carefully.
4. A stationary state is a hypothetical situation in which there is no economic growth. This means the income flow is uniform over time. Would there be an interest rate in a stationary state? Be sure to justify your answer.
5. Discuss some of the theoretical problems associated with using the market rate of interest as a guide to intertemporal resource allocation in the public sector.

ADDITIONAL READING

Alchian, Armen A. "The Rate of Interest, Fisher's Rate of Return over Cost, and Keynes' Internal Rate of Return," *American Economic Review, XLV* (December 1955), pp. 938–943.

Baumol, William J. "On the Social Rate of Discount," *American Economic Review, LVIII* (September 1968), pp. 788–802.

Dorfman, Robert (ed.), *Measuring Benefits of Government Investments*. Washington, D.C.: The Brookings Institution, 1965.

Eckstein, Otto. "Investment Criteria for Economic Development and the Theory of Intertemporal Welfare Economics," *Quarterly Journal of Economics, LXXI* (February 1957), pp. 56–85.

Fisher, Irving. *The Theory of Interest*. London: Macmillan and Co., Ltd., 1930.

Harcourt, G. C. "Some Cambridge Controversies in the Theory of Capital," *Journal of Economic Literature, VII* (June 1969), pp. 369–405.

Hirshleifer, J. "On the Theory of the Optimal Investment Decision," *Journal of Political Economy, LXVI* (August 1958), pp. 329–352.

Robinson, Joan. *An Essay on Marxian Economics,* 2d ed. New York: St. Martin's Press, 1966.

Usher, D. "Traditional Capital Theory," *Review of Economic Studies, XXXII* (April 1965), pp. 169–186.

11 Welfare Economics

We began our study of the theory of microeconomic policy with a discussion of criteria for evaluating the performance of an economic system. Such criteria are called **welfare criteria.** Since all economic systems, regardless of their social context, are faced with the problem of allocating scarce resources among competing uses, there are certain welfare criteria that are more or less universally accepted. But other standards of performance, particularly with respect to the optimal distribution of income, may vary considerably. In some countries, such as the United States, there is no performance criterion for income distribution that can truly be called a social goal.

In this chapter we use some of the tools of microeconomic analysis to reexamine the question of welfare criteria. In addition, we summarize our discussion of the role of public policy in optimal resource allocation.

11–1 EFFICIENCY CRITERIA

The New Welfare Economics of Pareto attempted to establish a universally valid criterion for evaluating economic performance. The Pareto criterion states that a change is an improvement if someone is made better off without making someone else worse off. If resources are allocated in such a way that no further change can make anyone better off without making someone else worse off, the allocation is said to be Pareto efficient or Pareto optimal.

It was noted that there are many Pareto efficient resource allocations associated with different distributions of welfare among households. Furthermore, the Pareto criterion does not state that a Pareto efficient allocation is better than one that is not Pareto efficient, if the two allocations involve a redistribution of welfare among individuals. However, for any inefficient allocation there is *some* better one that is Pareto efficient, since someone can be made better off without harming anyone else.

Although the Pareto criterion has universal applicability for

economies facing the need to allocate scarce resources among competing uses, the emphasis placed on efficiency relative to distributive considerations may vary considerably in different social contexts. Critics of the American economic system argue that too much emphasis has been placed on efficiency and too little on distribution. On the other hand, a case can be made for attempting to achieve efficiency once the desired income distribution has been attained. Many observers of socialist planning have criticized the inefficiency of the target output approach and have urged the use of a price system as a means of achieving allocative efficiency.

In this section we review the efficiency criteria in the light of the tools of analysis we have developed. Then we consider some distributive criteria.

Efficient Use of Productive Factors

Consider an economy consisting of two households, I and II, supplying factor services K and L. The economy produces two goods, food, F, and clothing, C. We will assume that both the goods and the factor services are homogeneous categories (for instance, the quality of the labor supplied by both households is the same.)

We will also assume that the production function is neoclassical, with strictly convex isoquants. This assumption is not necessary to the analysis, however. A finite activity production function is also acceptable as long as the isoquants are convex. However, assuming nonincreasing returns to scale and the absence of externalities is necessary to obtain our results.

Both K and L are inelastically supplied. Although this is not a necessary assumption for deriving the Pareto criteria, it will be convenient to assume that factor supplies are fixed.

Given these assumptions, we can derive the Pareto efficiency criteria using graphs. Figure 11–1 is an Edgeworth box diagram containing the isoquant maps for F and C. The width of the box is the amount of L supplied by the two households. The height is the amount of K supplied. The F isoquant map originates at the southwest corner of the box, and the C isoquant map is superimposed (upside-down), originating at the northeast corner.

Any point in the box represents a feasible production point. At P, for instance, FK_0 units of K and FL_0 units of L are used to produce F. CK_0 units of K and CL_0 units of L are used to produce C.

Although any point in the box is feasible, efficient production points occur only on the efficiency locus, where the F isoquants are tangent to C isoquants. Suppose resources are allocated to F and C at point Q, on isoquants F_2 and C_1. By reallocating resources to point P, production of C can be increased to C_2 with no reduction in F output. At P the C_2 isoquant is tangent to F_2. No further increase in C is possible without reducing the amount of F produced.

At any point on the efficiency locus no increase in C is possible without a reduction in F. Along the efficiency locus F isoquants are tangent to C isoquants,

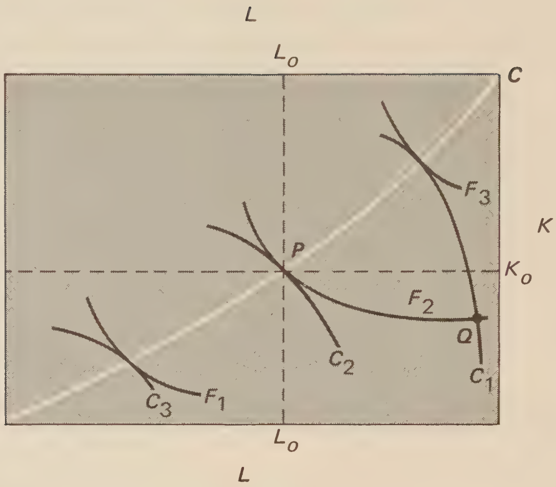


FIGURE 11–1
Efficient Use of Productive Factors

so that both sets of isoquants have the same slope there. Since the slope of an isoquant is the marginal rate of factor substitution in the production of a good, this implies that

$$MRS_{KL}^{F} = MRS_{KL}^{C}$$

Efficient allocation of resources between two goods occurs where the marginal rate of factor substitution is equal.

Optimal Output Mix

From the efficiency locus in Figure 11–1 a product transformation curve can be derived. The transformation curve shows the maximum amount of C that can be produced for a given amount of F. Each point on the transformation curve shown in Figure 11–2 is derived from the product indices along the efficiency locus in Figure 11–1. For example, point Z on the transformation curve at (F_2, C_2) corresponds to point P on the efficiency locus in Figure 11–1. Each point on the transformation curve represents an efficient output combination, since an increase in the output of one good always involves a reduction in output of the other. The slope of the transformation curve is the marginal rate of product transformation, that is,

$$-\frac{\Delta C}{\Delta F} = MRT_{CF}$$

The marginal rate of transformation is the amount of C that must be given up to obtain an additional unit of F when the economy is operating on the efficiency

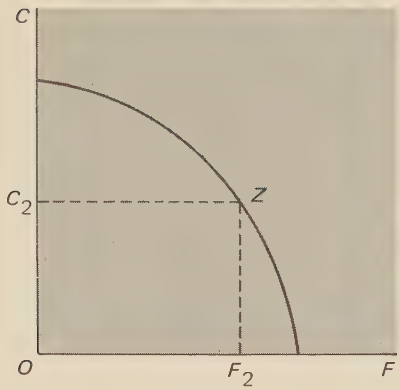


FIGURE 11–2
Efficient Production Possibilities

locus. MRT_{CF} is the opportunity cost of F in terms of C. Since the opportunity cost of F in terms of C is likely to rise as more F is produced (see page 143), the transformation curve will be convex to the origin.

The transformation curve provides us with a locus of output combinations that satisfy the efficiency criterion in production. The Pareto optimal output mix, that is, the optimal *point* on the transformation curve, will depend upon demand.

Consider a point, P, on the transformation curve. In Figure 11–3, P represents the dimensions of an Edgeworth box diagram containing the indifference maps of households I and II. C_0 is the total amount of C available, and F_0 is the total amount of F available. The slope of the transformation curve at P is the opportunity cost of F in terms of C. The indifference map of household I originates at the southwest corner of the box, and the indifference map of household II is superimposed (upside down), originating at the northeast corner.

Although any point in the box represents a feasible allocation of F and C between the households, trade will take place until the contract curve is reached, where the indifference curves of I are tangent to those of II. Suppose S represents the initial amounts of F and C held by the two households. Household I is at U_2^{I}, and II is at U_1^{II}. At S the amount of C that I is willing to give up for a unit of F is less than for II, that is,

$$\left(-\frac{\Delta C}{\Delta F}\right)^{\mathrm{I}} = MRS_{CF}^{\mathrm{I}} < MRS_{CP}^{\mathrm{II}} = \left(-\frac{\Delta C}{\Delta F}\right)^{\mathrm{II}}$$

Consequently, II could offer to purchase F from I at a price (in terms of C) in excess of MRS_{CF}^{I}. If II trades shrewdly, keeping the price just barely in excess of MRS_{CF}^{I}, it could move to point R. At R household II is better off than at S, while I is equally well off. However, II cannot improve its situation once R is reached except at the expense of I. Consequently, R is Pareto optimal.

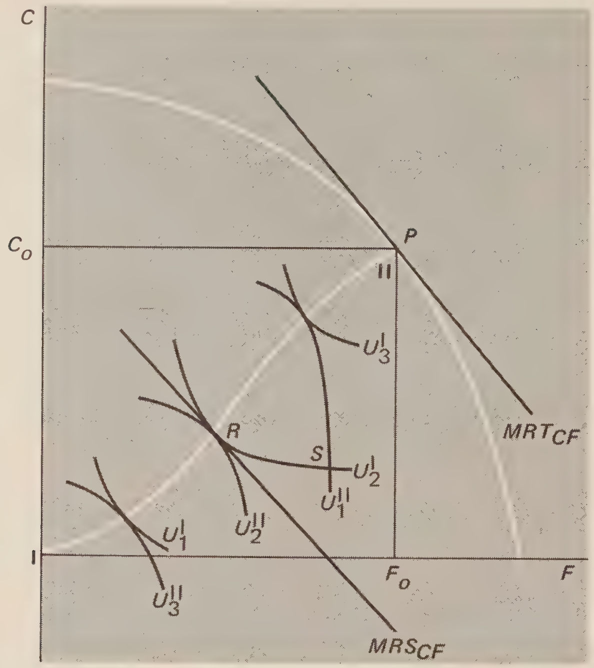


FIGURE 11–3
Optimal Output Mix

At R the U_2^{II} indifference curve is tangent to U_2^{I}. Along the contract curve indifference curves of I are tangent to those of II. Since the slope of an indifference curve is the marginal rate of product substitution, this implies

$$MRS^{I}_{CB} = MRS^{II}_{CB}$$

when goods allocation is Pareto optimal.

The contract curve represents the various allocations of C and F that are Pareto optimal in consumption. However, global efficiency requires that the rate at which consumers are willing to exchange goods be equal to their opportunity cost in production, so that the condition

$$MRS_{CB} = MRT_{CB}$$

must be satisfied.

Along the contract curve the slopes of the indifference curves vary. However, at some *point* on the contract curve, their slope, MRS_{CF}, will be equal to the slope of the transformation curve at P. In Figure 11–3 this occurs at R. Thus there is a single point in the box that represents the Pareto optimal distribution of C and F between the consumers, satisfying the efficiency condition

$$MRS^{I}_{CF} = MRS^{II}_{CF} = MRT_{CF}$$

At this point, I is at U_2^{I} and II is at U_2^{II}. Thus the point corresponds to a unique distribution of welfare (utility) between the households.

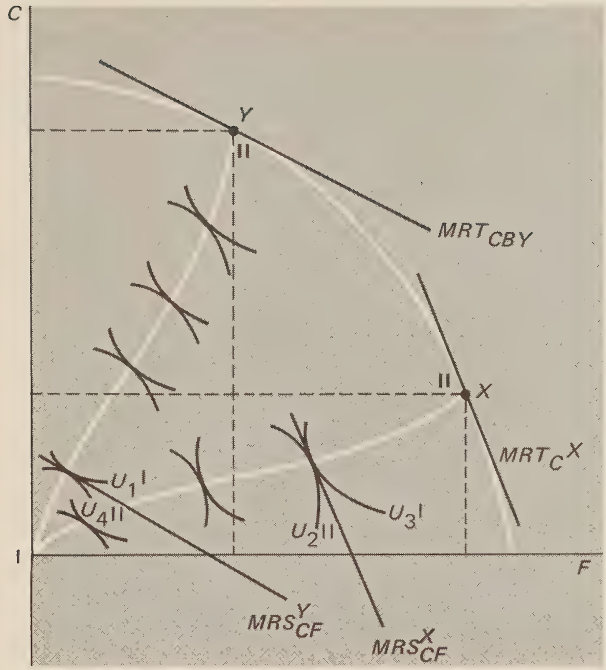


FIGURE 11–4
Alternative Output Mixes Associated with Different Distribution of Welfare

The Utility Trade-Off Frontier

If we repeat the process shown in Figure 11–3 for each point on the transformation curve, we will arrive at an equal number of Pareto efficient points satisfying the condition

$$MRS_{CF} = MRT_{CF}$$

Each will be associated with a different distribution of utility between households.

Figure 11–4 shows the Edgeworth box diagrams for two points, *X* and *Y*, on the transformation curve. At *X* relatively more *F* is produced. Efficient allocation of this amount of *C* and *F* occurs where

$$MRS^X_{CF} = MRT^X_{CF}$$

at U^{I}_3 and U^{II}_2. At *Y* relatively more *C* is produced and the distribution of welfare at the point where

$$MRS^Y_{CF} = MRT^Y_{CF}$$

changes to U^{I}_1 and U^{II}_4. At *Y*, II is better off and I is worse off than at *X*.

Since the optimal point within each box meets the Pareto criterion, to reach the optimal point in another box one household must be made worse off when the

other is better off. The welfare trade-off that takes place as the economy moves along the transformation curve is shown in Figure 11–5. The **utility trade-off frontier** is the locus of all combinations of ordinal utility[1] between households I and II that are attainable at different points on the transformation curve when the condition

$$MRS = MRT$$

is satisfied. All points on the utility trade-off frontier are Pareto optimal.

FIGURE 11–5
The Utility Trade-off Frontier

The Distribution of Welfare and Resource Allocation

Given any point on the utility trade-off frontier, we can determine the optimal product mix and the optimal utilization of factors for producing that mix. Suppose the initial distribution of welfare is at *A*. We can determine from Figure 11–4 the point on the transformation curve that corresponds to *A*. Given the product mix, we can from the corresponding point on the efficiency locus of Figure 11–1 determine how resources *K* and *L* should be allocated in production.

Clearly, the optimal allocation of resources depends upon the initial distribution of welfare between the households. At *B*, I is better off and II is worse off than at *A*. *B* corresponds to a different point on the transformation curve. In this example, production is likely to become more *F*-intensive as welfare is distributed

[1] An ordinal scale of preferences can be obtained from each household's preference map. Movement along an axis of the utility trade-off frontier simply implies a household is better or worse off. There is no attempt to measure utility in terms of the degree of psychic gratification associated with a reallocation.

in favor of I. In any event, a different product mix and a different allocation of K and L is associated with movements along the utility trade-off frontier.

The distribution of utility at any point on the utility trade-off frontier presupposes some distribution of factor ownership. Since factor payments are determined by the marginal rate of factor substitution in production, factor ownership may have to be redistributed to produce the distribution of income required to achieve that distribution of utility. On the other hand, if the distribution of factor ownership is predetermined, there is, in general, a single utility distribution associated with it. Thus we can select a desired distribution of utility and solve for the needed factor distribution. Alternatively, if the distribution of factor ownership among households is known, we can find the Pareto optimal distribution of utility associated with it.

The utility trade-off frontier is useful for demonstrating that there are many Pareto efficient resource allocations, each associated with a different distribution of welfare among the individuals in society. Although both A and B are Pareto optimal, a movement from A to B does not satisfy the Pareto criterion, since II is worse off after the move. Furthermore, even though point C is not Pareto optimal, a movement to A, which is Pareto optimal, will not satisfy the Pareto criterion, since I will be worse off after the move. However, a movement from C to B does meet the Pareto criterion for an improvement, since both households will be better off after the change. Thus a Pareto optimal allocation of resources is not necessarily preferred to a nonoptimal allocation, but if the allocation is nonoptimal, there will always be some Pareto optimal allocation of resources that represents an improvement.

11–2 DISTRIBUTIONAL CRITERIA

In searching for a criterion to evaluate the performance of an economic system we arrived at a utility trade-off frontier. The allocation of resources considered "best" depends ultimately on the way welfare is distributed among the individuals in society. Some economists feel that economics, as a social science, can make no normative judgments. But given the distribution of income (which, by and large, determines the distribution of welfare), we can determine the best use of economic resources.

Others argue that a welfare criterion that ignores distribution is a contradiction in terms. Resource allocation involves not only what is produced and how it is produced but *to whom* the product is distributed. These economists say that welfare economics has put too much emphasis on efficiency, on the grounds that it is a non-normative criterion not requiring value judgments, while ignoring the most fundamental welfare issue—distribution.

Although it is unlikely that Americans could ever agree about *the* optimal distribution of welfare among individuals, the question of distribution arises in nearly all economic decisions. Although monopoly is inefficient, measures to

regulate a monopolist to use marginal cost pricing will reduce his profits and thus reduce his welfare. A decision to build a highway in one place has different distributional implications than if the highway were located elsewhere. Although the Pareto criterion is useful for making an inefficient allocation efficient, it is not adequate for choosing between alternatives associated with different interpersonal distributions of welfare. And it is impossible to avoid the question of distributional criteria in most economic policy decisions.

Income Distribution and the Distribution of Welfare

Measures to redistribute the market-determined allocation of goods and services among individuals generally focus on the redistribution of income. While it is recognized that income is not the only source of economic welfare—leisure and the satisfaction or unpleasantness of one's job contribute to welfare too—policies that explicitly transfer income from some individuals to others can be assumed to transfer welfare in the same direction.

Other economic policies can redistribute welfare also. We have cited many instances in which the provision of public goods has redistributive consequences. Since, by definition, public goods are enjoyed equally by all, the provision of such goods will reduce the inequality in real income when it is financed by income taxes. Socialist countries with strong egalitarian ideologies generally provide public goods from taxes based on income, rather than direct income transfers, for redistributive purposes. Because income comes from the provision of productive factors by the household, income transfers may interfere with the optimal supply of these resources. This is particularly important when another economic goal is a high rate of growth, as has been the case in the Soviet Union and the countries of Eastern Europe.

Finally, another way in which welfare can be redistributed is through improved working conditions. Many low-status jobs are also unpleasant and affect the welfare of individuals holding them. On the other hand, high-status jobs are actually enjoyable and contribute to individuals' welfare. Policies that prohibit discrimination against blacks and women in these high-status jobs have definite redistributive implications with respect to welfare in addition to their effect on money income. The recent civil rights movements among blacks and women have been attempts to force white males to accept some of the unpleasant tasks traditionally relegated to these less privileged groups.

In the discussion that follows, we assume for simplicity that redistribution of welfare takes place through income transfers. Clearly, however, there are other ways that welfare can be redistributed.

The Kaldor-Hicks Criterion

Following the development of the New Welfare Economics by Pareto, economists in the 1930s attempted to derive distributive welfare criteria based on non-

normative considerations. That is, these criteria would not involve value judgments about the relative importance of individuals' welfare. The first attempt was by Kaldor and Hicks.[2]

Consider the utility trade-off frontier in Figure 11–6. Since all points on the frontier are Pareto optimal, movement along it from *B* to *C* requires a redistribution of welfare. Kaldor and Hicks assumed that such redistribution could be effected by lump sum transfers of money or wealth. Suppose the economy is at *A*. By the Pareto criterion, a move to *B* would not be an improvement, since it involves a loss for I. However, a move to *C* would be an improvement, since both I and II would be better off after the move. Kaldor argues that a move from *A* to *B* would also be an improvement, because at *B* we could redistribute wealth and move to *C*. Thus everyone is better off as a result of the change.

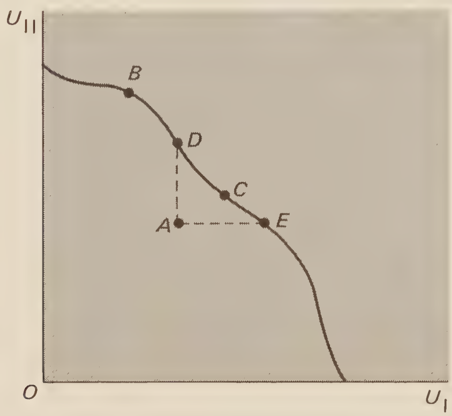


FIGURE 11–6
The Kaldor-Hicks Welfare Criterion

The Kaldor criterion can be viewed in another way. Suppose we move from *A* to *B*. Since *C* can be attained by a transfer of money, and since II is better off at *C* as well as at *B*, II should be willing to "bribe" I to move to *B*. The bribe must be large enough to move I to the right of *D* on the utility trade-off frontier. However, as long as the bribe does not move I past *E*, II will be willing to pay the bribe because he will be better off than at *A*.[3]

By the Kaldor-Hicks criterion, any point on the utility trade-off frontier is

[2] Nicholas Kaldor, "Welfare Propositions in Economics and Interpersonal Comparisons of Utility," *Economic Journal, XLIX* (September 1939), pp. 549–552; and J. R. Hicks, "The Foundations of Welfare Economics," *Economic Journal, XLIX* (December 1939), pp. 696–712.

[3] Tibor Scitovsky suggested the possibility that the losers might be able to bribe the gainers into *not* making the change. He suggested a "double criterion," that is, that *B* is preferred to *A* if the gainers could bribe the losers into accepting the change and in addition that the losers could not bribe the gainers *not* to make the change. See Tibor Scitovsky, "A Note on Welfare Propositions in Economics," *Review of Economic Studies, IX* (November 1941), pp. 77–78.

preferred to *any* point below it. If the economy is below the frontier, *some* point on the frontier meets the Pareto criterion for an improvement, and that point can be reached by an appropriate transfer of wealth.

Evaluation of the Kaldor-Hicks Criterion

The Kaldor-Hicks criterion contains the implicit assumption that II can compensate I for a loss in welfare by a direct money payment. Suppose II evaluates his gain at $500 and I evaluates his loss at $100. If II pays I an amount between $100 and $500, I will move into the *DE* segment of the utility trade-off frontier. With his increased wealth, I will purchase more goods and services, and a reallocation of resources will take place. Since the move from *A* results in a reallocation to the utility trade-off frontier between *D* and *E*, the move also meets the Pareto criterion for an improvement.

But suppose the compensation is not made. Even though II evaluates his gains at a greater money amount then I evaluates his losses, we cannot say that a move from *A* to *B* is an improvement. The marginal utility of income (or the marginal rate of substitution of goods for money) may be different for the two individuals. If I is very poor, the loss of welfare associated with a loss of $100 may be greater than the gain in welfare associated with a gain in $500 for II. Consequently, if the compensation is not paid, we cannot say the move from *A* to *B* is an improvement without making a value judgment about the relative importance of the welfare of the two individuals. If compensation is actually paid, however, the move would satisfy the Pareto criterion.

Social Welfare Functions

Most economists agree that any changes that do not satisfy the Pareto criterion must necessarily involve explicit value judgments about the best distribution of welfare among individuals in a society. Bergson, Samuelson, and others have suggested the need for a **social welfare function** that establishes alternative distributions of welfare equally satisfactory for the society as a whole.[4] The social welfare function would take the form of a set of social indifference curves between alternative welfare distributions, as shown in Figure 11–7. The point *B* at which the highest indifference curve is tangent to the utility trade-off frontier is the point of maximum social welfare, the **bliss point.** Notice that the bliss point is also Pareto optimal. If the economy is inside the utility trade-off frontier, a point on the frontier can be found that is an improvement. However, any point on the frontier is not necessarily preferred to any point inside it. In Figure 11–7, *D* is preferred to *C*. However, *B* is preferred to *D*.

Unlike the Pareto criterion, the social welfare function enables us to evaluate

---

[4] A. Bergson, "A Reformulation of Certain Aspects of Welfare Economics," *Quarterly Journal of Economics, LII* (February 1938), pp. 310–334; and Paul Samuelson, "Social Indifference Curves," *Quarterly Journal of Economics, LXX* (February 1956), pp. 1–22.

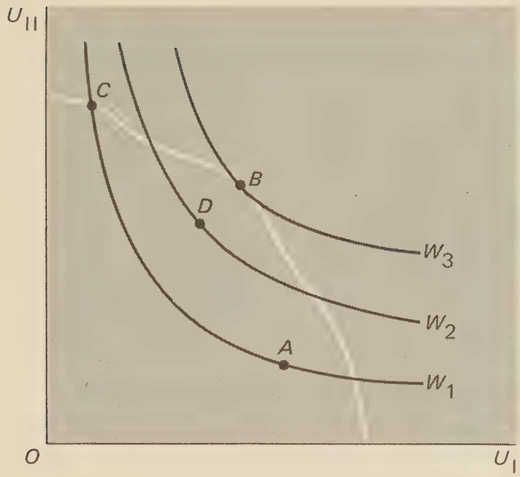


FIGURE 11–7
Use of a Social Welfare Function to Determine Optimal Resource Allocation

a move from *A* to *B* that involves a redistribution of welfare. However, the construction of the social welfare function involves making value judgments about the welfare of individuals.

The Theory of Social Choice

Formulation of a social welfare function requires certain explicit choices concerning the best distribution of welfare among members of a society. Such choices can be made by a dictator, a representative assembly, or a democratic process in which all members of society have an equal vote.[5]

Kenneth Arrow has shown that social preferences resulting from voting may be inconsistent, even if individual preferences are consistent.[6] That is, for an individual if

A is preferred to *B* is preferred to *C*

then

A must be preferred to *C*

However, for a group, if

A is preferred to *B* is preferred to *C*

[5] Giving everyone an equal vote is a value judgment in itself. Does each household get a vote, each person, each adult? Why should everyone get an equal vote if their capacities to produce or to enjoy goods are different?

[6] Kenneth Arrow, *Social Choice and Individual Values* (New York: Wiley, 1951).

then

$$C \text{ may be preferred to } A$$

Furthermore, Arrow has shown that social choices may be different from those of individuals, and social rankings may be different from private rankings.

Suppose three individuals, I, II, and III, are asked to rank three projects, *A*, *B*, and *C*. They are asked to vote by writing 3 for their first choice, 2 for their second choice, and 1 for their first choice. Their votes are as follows:

	A	*B*	*C*
I	3	2	1
II	1	3	2
III	2	1	3

Each individual has consistent preferences. Two individuals, I and III, prefer *A* to *B*. Two, I and II, prefer *B* to *C*. Yet the majority prefers *C* to *A*. The preference of the majority is inconsistent with those of the individuals composing it.

Arrow's work indicates that the use of the democratic process of voting for formulating a social welfare function may produce contradictory welfare criteria. However, welfare economics has not come up with an alternative to a social welfare function involving explicit value judgments for providing distributional criteria. In a society with a universally accepted ideology related to distribution the formulation of explicit value criteria is no great problem. However, in a country like the United States, with its diverse opinions and interest groups, it is no wonder that economists have traditionally sidestepped the issue of distribution.

Evaluation of Modern Welfare Economics

The current state of thinking about welfare economics among professional economists is divided. Some economists like Kenneth Boulding argue that welfare economics can tell us nothing about optimal resource allocation, since it has failed to provide a distributive criterion.[7]

Samuelson, on the other hand, argues that since Pareto efficiency is a necessary (although not sufficient) condition for reaching the social optimum, we can say quite a lot about optimal resource allocation. We can show, for instance, that point rationing is better than fixed quantity rationing, or that income subsidies are better than subsidies in kind. While admitting that there are many decisions for which welfare criteria are not available, he says that welfare economics is a useful way to examine the consequences of various value judgments whether or not they are shared by the theorist.[8]

---

[7] See Kenneth Boulding, "Economics as a Moral Science," *American Economic Review, LIX* (March 1969), pp. 1–12.

[8] Samuelson's view of the methodological implications of welfare economics can be found throughout his work. One source is *Foundations of Economic Analysis* (Cambridge, Mass.: Harvard University Press, 1947), chapter 8.

11-3 SPECIAL PROBLEMS IN WELFARE ECONOMICS

In addition to the problem of devising distributional criteria, there are some basic allocative problems that microeconomic theory has not been able to resolve satisfactorily. Although we have already discussed these problems, we briefly summarize their implications in this section.

Technical Versus Allocative Efficiency

The existence of economies of scale in production due to factor indivisibilities, in advertising, and in undertaking research and development activities for technological change implies that the most efficient size of firms in many industries may be so large that a firm's output decisions will have a significant impact on market supply and price. In such a market firms will produce where

$$MR = MC$$

and since

$$p > MR$$

this implies

$$p > MC$$

In the absence of externalities, the condition for allocative efficiency is

$$p = MC$$

Thus firms experiencing economies of scale underproduce relative to the social optimum.

If firms are maximizing profits, the only way to force them to adopt marginal cost pricing is to divide them into units so small that each has a negligible impact on market supply and price. However, if there are legitimate economies of scale, unit costs will rise and technological efficiency will be lost. In some cases, costs may increase so much that output actually falls below the initial level.

The Lange solution, which fixes the price at which a firm can sell its output regardless of the scale of operation, lacks a mechanism for ensuring the market will be cleared. If there is excess supply or demand at the established prices, Pareto efficiency will not be achieved, even though all firms are setting price equal to marginal cost. If a central board continually readjusts prices in response to excess supply or demand, firms will act as if they were monopolists, knowing that if they restrict output and produce excess demand, prices will eventually rise. Consequently, they can estimate marginal revenue and will set $MR = MC$, provided they have any incentive at all to maximize profits.

There is no simple answer to this dilemma. Regulation of public utilities in the United States presumably aims to set price equal to marginal cost. However, there is no way to determine the optimal scale of output. When there is no profit incentive, and the emphasis is on the relation of price to cost rather than to the

services provided, the consumer does not necessarily benefit. Growing pressure for the United States Post Office to be run as a private corporation, complete with profit incentives, is evidence that regulation is not necessarily the solution.

Externalities

The existence of externalities is the perennial allocative problem for economic theory. Most discussions begin with the phrase, "assuming no externalities. . . ." However, externalities are less a theoretical than a practical problem. In chapter 3 it was noted that an externality (in production) involves a discrepancy between the marginal value product of resources from the viewpoint of the firm (or individual) making allocative decisions and the marginal value product of resources from the viewpoint of society. Efficient allocation of resources occurs when resources are paid their marginal *social* value product, MVP_s. However, firms in both the market model and in a Lange economy employ resources where the resource price equals the marginal *private* value product, MVP_p. Theoretically, government can design a tax or subsidy that will equate MVP_s with MVP_p.

The difficulty arises in determining the discrepancy between MVP_s and MVP_p. As noted, the discrepancy depends upon the rate at which the externally produced good (or bad) is generated per unit of private output. Furthermore, the social value or cost of the externality must be determined. Since many externalities—air pollution, beauty, noise—are public goods (or bads) there is no market price available and shadow prices must be imputed.

Once the invisible hand of the market or the automatic adjustment of firms to the prices set by a central planning board in the Lange model breaks down for producing efficient resource allocation, explicit knowledge of production relations and complicated valuation procedures are required. Consider the problem of assessing the social cost of an automobile. Exactly how much pollution, traffic congestion, inconvenience to pedestrians, and so on, can be attributed to a single automobile? And even if this could be determined, what is the social cost of these external effects?

Intertemporal Welfare Considerations

In chapter 10 we considered the Pareto criterion for the distribution of resources over time. Allocation of resources for future consumption is efficient when the social rate of time preference is equal to the internal rate of return on capital. In an economy where wealth is privately owned, resources are released for future consumption at a rate that reflects the rate of time preference of the current adult generation. While this may be appropriate for determining the optimal expenditure pattern of a household for goods that yield benefits over a period shorter than the life expectancy of adults—automobiles, washing machines, even residential housing—it is not necessarily the correct rate for determining the optimal level of social investment.

Consider a project designed to control the rate of air pollution. If it is begun immediately, the project will cost $10 billion per year. If it is postponed, the cost of reducing pollution to the same level will increase rapidly. Although the annual social cost of removable air pollution at the present time may not be as great as $10 billion, if present trends continue, the social cost is likely to exceed that figure in twenty-five years. Suppose it is estimated that in thirty years the social cost of removable air pollution will be $110 billion per year, $100 billion in excess of the cost of eliminating it if the project were begun now. At an interest rate of 6 per cent, the present value of $100 billion in thirty years is only $17.4 billion.[9] At a zero rate of interest, of course, the present value is $100 billion. Clearly, the evaluation by the current generation of the social cost of pollution for future generations is extremely sensitive to the interest rate selected to compute the present value of the costs. At 6 per cent the annual social costs after thirty years barely exceed the annual outlay required for the program, which must begin immediately. At a lower rate of interest, the social costs of pollution after thirty years are more than five times the annual outlay for the pollution abatement program. Thus, after thirty-six years the reduction in social costs would compensate for the expenditure on the project.

Since the current market rate of interest reflects the time preference of the present generation, it is generally unsuitable for discounting the returns to social investment projects. However, the selection of any other interest rate for this purpose is also arbitrary. The choice essentially involves a value judgment about the relative interests of the present and future generations. The lower the interest rate used for calculating present value, the greater the benefit to future generations relative to the present generation. Thus the question of the optimal allocation of resources to the future emphasizes again the need for formulating distributive criteria.

11-4 THE THEORY OF MICROECONOMIC POLICY: A SUMMING UP

Both the noeclassical market model of perfect competition and the Lange model of decentralized socialism were attempts to show that households' and firms' reactions to a set of fixed prices will produce an efficient allocation of resources. Although neither model deals systematically with the question of distribution, both imply that redistribution of income can be effected through appropriate wealth or income transfers. Once a tax scheme is designed to transfer wealth or income and an optimal income distribution achieved, then the market system

[9] The present value of $100 billion after thirty years is

$$PV = \frac{\$100 \text{ billion}}{(1+i)^{30}}$$

where i is the interest rate.

or the Lange system could be relied on to produce the optimal allocation of resources.

It was noted that the market model of perfect competition could never exist in practice because of economies of scale. Bator calls this "failure of structure."[10] But even if we could avoid the problem of monopoly (as Lange attempted to do), we would still have the problem of the discrepancy between private and social values, which renders private decisions inconsistent with socially desirable ones. Pollution, noise, inadequate consideration of future generations will always exist as long as firms take only their own costs and benefits into account when deciding how much to produce and invest.

We have discussed the implications of these problems for microeconomic policy. But there is still much work to be done in this area. For many years, the laissez faire tradition obscured the debate over relevant policy issues. Economists who recommended government interference in the market mechanism were thought to be undermining the capitalist system. But just as government regulation of aggregate demand is necessary for the long-run survival of capitalism, as Keynes pointed out, microeconomic policy is important in enabling the free enterprise system to withstand the censure of its critics.

Socialist solutions do not seem any more promising, at least with respect to the efficiency issue. In fact, socialist systems seem to be falling back on the market mechanism as the best way to allocate resources. This trend is also apparent within large enterprises and in the public sector in capitalist countries. The use of shadow prices for planning when market prices are not available, product competition within firms, decentralization of decision making are all characteristic of this phenomenon.

Although we have suggested areas in which public policy can be used to improve the efficiency of a market system, we have left open the question of the appropriate scope of these policies. There is no answer to this question upon which we can all agree. The costs of administration must be weighed against the improvements in efficiency. And since most improvements in efficiency also involve redistribution of welfare, the actual benefit derived is open to subjective evaluation.

QUESTIONS FOR STUDY AND REVIEW

1. If completely free international trade implies maximum economic welfare, does it necessarily follow that *any* increase in trade would improve welfare? On the other hand, can you show conclusively that *some* trade is always better than no trade at all? Illustrate your answer with a graph or with a numerical example.
2. Discuss the concept of Pareto optimality as a welfare criterion, including both its

[10] See Francis M. Bator, "The Anatomy of Market Failure," *Quarterly Journal of Economies, LXXII* (August 1958), p. 355.

strengths and limitations. How is Pareto optimality related to the concept of a social welfare function? What are some of the problems involved in defining a social welfare function?

3. "In our economy, by and large, the future can be left to take care of itself. There is no need to lower artificially the social rate of discount in order to increase further the wealth of future generations. The rate of interest should presumably then be set by the market and the needs of public policy . . . and no attempt should be made to subsidize the future by artificial reductions in discount rates designed only for that purpose." Discuss.

4. Describe and compare the welfare implications of the following alternative schemes for running a country's radio-television network.* In all cases, the managing agents are assumed free of government control over program content. Assume that the country is small enough to constitute a single "listening area," so that the problem of geographical distribution can be ignored.
   a) A national commission distributes a limited number of station licenses by the criterion of "applicant's merit." The station owner is a profit maximizer. The operation is financed through advertising, the station selling "spots" or "sponsorships" tied to specific programs.
   b) The situation is as in (a), except that licenses are auctioned off to the highest bidder.
   c) The situation is as in (a), except that no advertising is allowed. Instead, meters are installed on all receiving sets, showing the total time over which each station was received. Set owners are charged a tax proportional to total reception time, and the proceeds are distributed to stations in proportion to their share of the aggregate reception time.
   d) Station managers are salaried employees of the national commission and decide on programs on the basis of a set of general rules issued by the commission. A flat tax is levied on all sets, and the proceeds are shared equally among all stations. Each station must operate within the corresponding budget.

5. Discuss and evaluate the present state of welfare economics.

ADDITIONAL READING

Arrow, Kenneth J. *Social Choice and Individual Values*. New York: Wiley, 1951.

Bator, Francis M. "The Simple Analytics of Welfare Maximization," *American Economic Review, XLVII* (March 1957), pp. 22–59.

Baumol, William J. *Welfare Economics and the Theory of the State*. London: Longmans, Green, 1952.

Hicks, J. R. "The Foundations of Welfare Economics," *Economic Journal, XLIX* (December 1939), pp. 696–712.

Kaldor, Nicholas. "Welfare Propositions in Economics and Interpersonal Comparisons of Utility," *Economic Journal, XLIX* (September 1939), pp. 549–552.

* This problem is from an examination written by André Danière at Harvard University in 1964.

Little, I. M. D. *A Critique of Welfare Economics.* Oxford: Oxford University Press, 1950.

Mishan, E. J. "A Survey of Welfare Economics, 1939–1959," *Economic Journal, LXX* (June 1960), pp. 197–256:

Samuelson, Paul. "Evaluation of Real National Income," *Oxford Economic Papers, II* (January 1950), pp. 1–29.

Scitovsky, Tibor. "Reconsideration of the Theory of Tariffs," *Review of Economic Studies, IX* (1941–42), pp. 89–110.

INDEX

Definitions of terms appear on pages indicated in **bold face** type.

1 2 3 4 5 6 7 8 9 10

0-669-83170-0